Scientometrics, Vol. 30, No. 1 (1994) 213-227

DOUBLE EXPONENTIAL MODELS FOR FIRST-CITATION PROCESSES*

R. ROUSSEAU

Informatie- en Bibliotheekwetenschap, University of Antwerp (UIA), Universiteitsplein 1, 2610 Wilrijk (Belgium)
and
Katholieke Industriële Hogeschool West-Vlaanderen, Zeedijk 101, 8400 Oostende (Belgium)

(Received January 18, 1994)

The purpose of this article is to find a model for the first-citation or response distribution. Starting from plausible assumptions, we derive differential equations, whose solutions yield the requested functions. In fact, we propose two different double exponential distributions as candidates to describe the first-citation process. We found that some real data are best fitted by the first of these models and other by the second. We further note that Gompertz' curve plays an important role in this second model. These models can be used to predict the total number of articles in a fixed group that will ever be cited. We conclude that further research is needed to find out when one of the two models is more appropriate than the other.

Introduction

Citation analysis can be described as that subfield of informetrics which studies the flow of information using reference lists. It has an important branch that concentrates on descriptive techniques[1] (lists of most-cited articles, authors, institutes, etc.), but also more advanced statistical techniques and mathematical models play a prominent role. For a review on citation analysis we refer the reader to Part III of our book.[2] We would further like to draw the reader's attention to Peritz' note[3] on the objectives of citation analysis and her appeal for a substantial methodological input (in the field) from statisticians and informetricians.

Whenever a group of articles is published this event can be considered as the introduction of a stimulus in Popper's World III (see Refs 4, 5 for a discussion of information space and its relation to Popper's World III). Then citations, as symbols for 'use', can be interpreted as responses to this stimulus. Total response occurs when every article of the group has been cited at least once. Normally, this will not

*Paper presented at the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics in Berlin (Germany), September 11-15, 1993.

Elsevier, Amsterdam – Oxford – New York – Tokyo
Akadémiai Kiadó, Budapest

occur as many articles, even those published in prestigious journals remain uncited.[6,7] Taking 'use' as a yes/no situation, we consider a total response as built up from partial responses, namely citations (at least one, but the number of citations will not play a role in this investigation as citation is considered as a yes/no situation) of a particular article of the group under consideration. In this article, we are interested in the gradual building up of total response. By 'group' we mean a set of publications published in the same journal, or in the same language, by the same author, or members of the same research institute.

From a model-theoretic point of view our aim is to find the 'first-citation' or response time distribution for journal articles. We will show that a well-known double exponential model, which has been used in many other fields, e.g., in sociology[8] can be used here, too. Response time has been studied before, e.g., by *Moed* and *Van Raan*[9] (where it has been called 'citation delay') and by *Schubert* and *Glänzel*[10] who propose the mean response time (MRT) as a new journal citation indicator of immediacy. (For the exact definition of MRT and its relation to other citation measures we refer the reader to Ref. 10.)

We finally mention that this article is a shortened version of a report with the same title.[11]

Construction of a simple model

Consider a fixed group of N articles. Let $C(t)$ be the cumulative number of articles cited at least once by a fixed group of sources (e.g., a particular journal or all ISI journals or some particular group of scientists) over a period of length t. We assume that the change in $C(t)$ is proportional to the number of uncited articles, with a proportionality factor denoted as q. The rationale for this assumption is that the most interesting articles are picked up at a fast rate, while the lesser interesting (or more difficult) articles are accepted at a slower and slower pace.

The constant q can be interpreted as a conversion factor: it describes the rate at which articles shift from the uncited group to the cited one. Of course, q depends on the group of articles under consideration. This leads to the following linear differential equation.

$$\frac{\mathrm{d}\,C(t)}{\mathrm{d}t} = q\,(N - C(t)), \qquad 0 < q \tag{1}$$

Using $R(t) = C(t)/N$, i.e., the cumulative, relative number of cited articles up to time t, equation (1) becomes:

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = q\,(1 - R(t)) \tag{2}$$

Solving this differential equation by separation of variables leads to the following equalities:

$$\int_{R(0)}^{R(t)} \frac{\mathrm{d}R(t')}{1 - R(t')} = \int_{0}^{t} q\,\mathrm{d}t' \tag{3}$$

where we assume that $R(t) \neq 1$ for every finite time t.

Integrating equation (3) gives:

$$\frac{1 - R(0)}{1 - R(t)} = e^{qt} \tag{4}$$

or,

$$R(t) = 1 - k\,e^{-qt}, \qquad \text{where } k = 1 - R(0). \tag{5}$$

Equation (5) is actually Ware's growth model[12] for relative values. If $R(0) = 0$, then $k = 1$. However, we will assume that the citation phenomenon begins at time 0, where 0 does not necessarily coincide with the time of publication, in which case $R(0) \neq 0$, and k is a constant to be estimated from the data. If equation (5) would describe the first-citation process, how many articles would be cited ultimately. In other words, what happens in the limit, for $t \to \infty$?

$$\lim_{t \to \infty} R(t) = \lim_{t \to \infty} (1 - k\,e^{-qt}) = 1 \tag{6}$$

Hence, it follows from the assumptions of this simple model that eventually all articles are cited. However, it is well-known that in practice, many articles remain uncited. Consequently, we will alter the simple model to incorporate this fact.

A double exponential model

In the above model we have assumed that the conversion factor q is a constant. However, it seems more reasonable to assume a time-dependence. Indeed, the older an article is the less likely it is to be cited, and this for several reasons. First, scientists tend to cite the more recent articles (because these are best remembered, or, perhaps, scientists want to show that they are aware of the recent literature); second, older articles sometimes contain facts that are not correct anymore, or are otherwise superseded by more recent findings.

Now, we assume that q also depends on t. We will model this time-dependence by a negative exponential distribution:

$$q(t) = A\, e^{-at} \;; \quad A > 0, \quad a \geq 0 \tag{7}$$

Substituting (7) into (2) yields:

$$\frac{d R(t)}{dt} = A\, e^{-at}\,(1 - R(t)) \tag{8}$$

Also Eq. (8) is a first order linear differential equation that can be solved by the technique of separation of variables. This yields:

$$\int_{R(0)}^{R(t)} \frac{d R(t')}{1 - R(t')} = \int_0^t A\, e^{-at'}\, dt' \tag{9}$$

(where again we assume that $R(t) \neq 1$),

$$< = = = > \qquad \frac{1 - R(0)}{1 - R(t)} = e^{(A/a)(1-e^{-at})} \quad (\text{if } a \neq 0) \tag{10}$$

Putting $k = 1 - R(0)$, $0 < k \leq 1$; and $e^{-A/a} = b$, hence $A = -a \ln(b)$; $0 \leq b < 1$, equation (10) becomes:

$$R(t) = 1 - kb^{(1-e^{-at})} \tag{11}$$

Now

$$\lim_{t \to \infty} R(t) = 1 - kb \leq 1 \tag{12}$$

which is more realistic. Analyzing equation (11) shows that $R(t)$ is a function that takes values between 0 and 1, increases and is always concave for the stated values of the parameters (i.p. $a > 0$). If $a = 0$, equation (8) coincides with equation (2). $R(0)$ represents the beginning of the citation phenomenon: time 0 is not necessarily the time of publication.

A check of the model

A mathematical model has no practical value unless it fits real data. To check this double exponential model for first-citation data we will apply it to two sets of data. The first one consists of references in the Russian scientific literature to Russian language library science periodicals, as published by Motylev.[13] The second one consists of all cita-tions to articles published in the *Journal of the American Chemical Society* (JACS) in 1975 (i.e., Vol. 97) in this same journal. We have collected the latter data ourselves.

Data for the Russian-language library science and bibliography periodicals[13] are given in Table 1.

Table 1a

Articles published in 1960: total 549

A: Year (i);
B: Number of articles, cited for the first time during the i-th year

A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
B	10	33	18	17	14	4	0	10	7	7	8	9	4	13	6	6

Table 1b

Articles published in 1965: total 617

A	1	2	3	4	5	6	7	8	9	10	11	12
B	15	29	28	51	18	15	9	8	15	10	12	2

For articles published in 1960 we find the following parameter values (using Marquardt's algorithm[14] for nonlinear least squares):

$k = 0.959$
$b = 0.652$
$a = 0.081$, hence $A = 0.035$

where the coefficient of determination $R^2 = 0.9744$ and the residual sum of squares is 0.0024. This yields the following equation for $R(t)$:

$$R(t) = 1 - 0.959 \times 0.652^{(1-e^{-0.081\,t})} \qquad (13)$$

The number $(R^2 \times 100)\%$ represents the percentage of total variability explained by the model. The citation phenomenon begins at time 1. Taking into account that data are fairly irregular this fit is excellent, see Fig.1. A Kolmogorov-Smirnov test shows that the fit is accepted even at the 10% level (so certainly at the 5% and the 1% level).

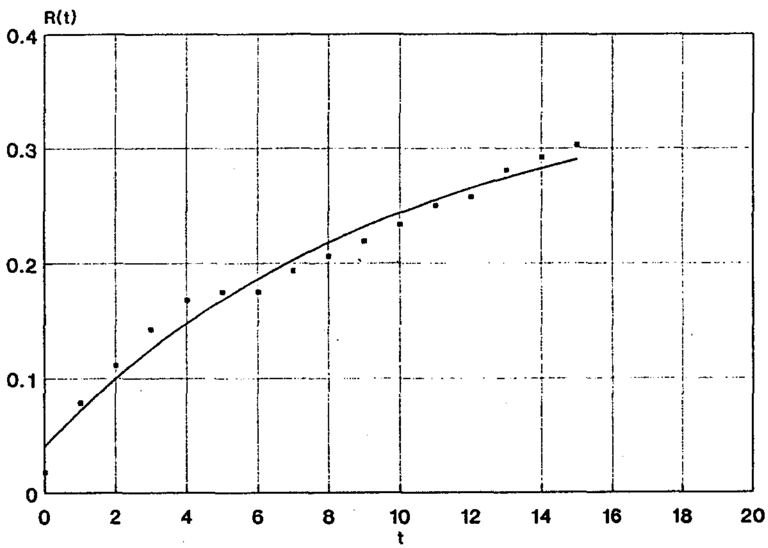

Fig. 1. MOTYLEV 1960 data. Best fitting curve using the first double exponential model

Similarly, for articles published in 1965 we find:

$k = 0.985$

$b = 0.608$

$a = 0.154$, hence $A = 0.077$,

where the coefficient of determination $R^2 = 0.989$ and the residual sum of squares is 0.0013. Hence:

$$R(t) = 1 - 0.985 \times 0.608^{(1-e^{-0.154\,t})} \tag{14}$$

This fit is accepted by a Kolmogorov-Smirnov test at the 10% level. This result is even slightly better than the one for 1960 publications. The predicted percentage of articles that will ever be cited (by other Russian publications) is resp. 37.5% (for 1960 publications) and 40% (for 1965 publications).

Next, we consider the JACS data. These data consist of response times of all articles published in JACS during the year 1975 (26 issues). As this journal is published biweekly the time unit used in these investigations is two weeks. The JACS publishes two kinds of articles: 'leading' or full length articles and communications. Observations for both groups are given separately in Table 2. We have collected data for the first four years after publication, assuming that by that time the majority of citable articles of this journal are indeed cited. After 4 years the percentages cited are: 68.6% for leading articles, 66.3% for communications and 67.6% for all JACS articles.

Table 2

First-citations of JACS articles in JACS

- A: number of issues since publication
- B: number of cited leading articles
- C: number of cited communications
- D: sum of B and C

A	B	C	D	A	B	C	D
0	4	3	7	53	4	3	7
1	0	0	0	54	3	3	6
2	0	0	0	55	5	5	10
3	1	0	1	56	5	3	8
4	2	3+1	5+1	57	6	3	9
5	5	9	14	58	6	2	8

(Table 2 cont.)

6	2	6	8	59	5	5	10
7	3	5	8	60	1	3	4
8	4	3	7	61	3	4	7
9	3	5	8	62	7	2	9
10	7	7	14	63	6	4	10
1	8	2	10	64	8	2	10
2	11	5	16	65	13	0	13
3	7	3	10	66	1	2	3
14	12	14	26	67	4	4	8
15	12+2	11	23+2	68	4	5	9
16	11	4+1	15+1	69	4	4	8
17	6	10	16	70	5	4	9
18	15	10+1	25+1	71	4	0	4
19	13	8	21	72	6	0	6
20	15	11	26	73	3	2	5
21	13	12	25	74	7	3	10
22	20	7	27	75	5	0	5
23	13	10	23	76	3	2	5
24	15	8	23	77	2	1	3
25	11	19	30	78	3	3	6
26	15	16	31	79	5	1	6
27	19	13	32	80	1	6	7
28	18	15	33	81	3	2	5
29	10	12	22	82	3	2	5
30	14	6	20	83	2	4	6
31	20	12	32	84	0	3	3
32	17	6	23	85	1	1	2
33	18	10	28	86	3	1	4
34	10	12	22	87	1	0	1
35	23	11	34	88	0	3	3
36	12	2	14	89	4	0	4
37	14	10	24	90	1	0	1
38	9	14	23	91	3	3	6
39	11	0	11	92	1	3	4
40	7	7	14	93	2	3	5
41	11	10	21	94	2	1	3
42	13	11	24	95	2	0	2
43	14	7	21	96	2	2	4
44	11	5	16	97	0	2	2
45	7	7	14	98	2	3	5
46	9	5	14	99	5	1	6
47	4	7	11	100	2	1	3
48	6	5	11	101	4	0	4
49	7	0	7	102	0	1	1
50	4	9	13	103	4	0	4
51	6	8	14	104	2	0	2
52	9	5	14				
					707	512	1219

Published articles: 1031 leading articles, 772 communications, hence 1803 in total.

As data for this set were collected by ourselves, we tried to obtain as pure a set as possible. For instance, we counted only citations by other scientists. The reason for omitting self-citations is that the citation processes for self-citations and for citations by other scientists differ. We also omitted citations in the same issue as the one in which the article is published. Again, here, citation is not a reaction to an existing publication, or even to a preliminary draft, but there might be another phenomenon at work: e.g., a direct intervention of the journal editor. For those articles which are in this case we mentioned their second citation (the +'s in Table 2; note that for unknown reasons some of these articles are not cited a second time!?). It is this second citation that is used in the calculation of $R(t)$.

Results of the fitting exercise for the JACS data:

Leading articles:

$k = 1.014$
$b = 120.68$
$a = -0.0026$ $R^2 = 0.97$

Communications:

$k = 0.985$
$b = 393.4$
$a = -0.0020$ $R^2 = 0.96$

All articles:

$k = 1.003$
$b = 322.7$
$a = -0.0021$ $R^2 = 0.97$

Contrary to the results for the Russian data our double exponential model does not fit the data very well. Although the R^2-values are high (showing that R^2 is not an appropriate measure for goodness-of-fit), these results make little sense: the a-values should be positive, but we have obtained (small) negative values. Indeed, as these a-values are very small, we could as well have taken a linear regression model. These observations are confirmed by a Kolmogorov-Smirnov test: fits are rejected even at the 1% level.

The reason for this bad fit is obvious from Fig. 2. The observed curve is S-shaped – has an inflection point – but the model cannot incorporate this. In the next section we will introduce another model correcting this flaw.

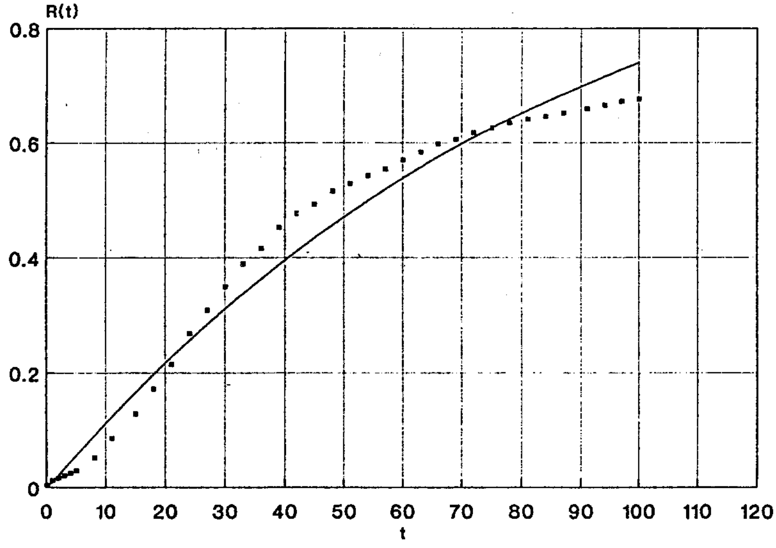

Fig. 2. JACS data. Best fitting curve using the first double exponential model (only selected points are shown; fitting has been based on all points)

A double exponential model, leading to an S-shaped curve

To solve the problem that the model introduced in the previous section leads to a curve without inflection points, we introduce a factor $R(t)$ in differential equation (8). This new equation [Eq. (15)] is now a generalisation of the growth model leading to a logistic curve and coincides completely with the differential equation used by *Hernes*[8] to model the (first) entry into marriage:

$$\frac{\mathrm{d}\,R(t)}{\mathrm{d}t} = A\,\mathrm{e}^{-at}\,(1 - R(t))R(t) \qquad (15)$$

There is, however, an important difference between this model for the first-citation process and population growth models. When studying growth, it is natural to take as a first assumption that the change in growth depends on the population: the more productive members in the population there are, the larger the rate of growth. Without any restricting factors this assumption leads to exponential growth and it is the merit of Verhulst[15] that he has introduced in the growth model a factor corresponding to our $(N - C(t))$, leading to the logistic curve.

Here, in our first-citation model the factor $R(t)$ is introduced last, as a correction factor, to account for the S-shape at the beginning of the observed curve. The introduction of this factor can be defended by observing that the more articles of a coherent group (e.g. which have appeared in the same journal) are cited, the more likely it becomes that articles of this group come to the attention of scientists, are read attentively, and used and referred to in later research.

Equation (15) is not linear anymore, yet it can still be solved by separation of the variables. Integrating (15) between $R(0)$ and $R(t)$ gives:

$$\int_{R(0)}^{R(t)} \frac{\mathrm{d}R(t')}{R(t')(1-R(t'))} = \int_0^t A\,e^{-at'}\,\mathrm{d}t' \quad \text{with } R(0) \neq 0$$

or
$$-\ln\left(\frac{1-R(t')}{R(t')}\right)\Bigg|_{R(0)}^{R(t)} = \frac{A}{a}\,e^{-as}\Bigg|_0^t$$

(where we assume that $a \neq 0$)

or
$$= \ln\left(\frac{1-R(0)}{R(0)}\right) - \ln\left(\frac{1-R(t)}{R(t)}\right) = \frac{A}{a}\,(1-e^{-at}).$$

This can be rewritten as:

$$\frac{R(t)}{1-R(t)} = \frac{R(0)}{1-R(0)}\; e^{A/a.(1-e^{-at})}$$

Putting $Q_0 = R(0)/(1-R(0))$, $b = e^{-A/a}$ and $M = Q_0/b$, we obtain the following double exponential model for $R(t)$:

$$\frac{R(t)}{1-R(t)} = M\,b^{e^{-at}} \tag{16}$$

or

$$R(t) = \frac{1}{1 + \dfrac{1}{M\,b^{e^{-at}}}} \tag{17}$$

Note that $0 < Q_0 < +\infty$; $0 < b < 1$; $0 < M < +\infty$. As R(0) ≠ 0 (and is in fact equal to $(1 + (Mb)^{-1})^{-1}$), $R(0)$ represents the beginning of the citation phenomenon. Again, time 0 is not (necessarily) the moment of publication. The right-hand side of (16) is Gompertz' curve, well-known from investigations in other fields.[16,17] This curve has also been studied by *Egghe* and *Rao*[18] to describe the growth of the social science and humanities literature as shown in on-line databases. Hence, although the first-citation distribution is not given by Gompertz' equation, a simple function of $R(t)$ is.

Note further that

$$\lim_{t \to \infty} R(t) = M/(1+M) < 1 \tag{18}$$

showing that in this model too, not all articles are eventually cited.

Fitting the JACS data

For Gompertz' model [equation (16)], we have obtained the following values for the parameters of best fitting curves:

Leading articles: $M = 2.169$; $b = 0.0060$; $a = 0.04167$; with $R^2 = 0.998$ and a residual sum of squares equal to 0.123;

Communications: $M = 1.983$; $b = 0.0097$; $a = 0.04210$; with $R^2 = 0.997$ and a residual sum of squares equal to 0.120;

All articles: $M = 2.087$; $b = 0.0074$; $a = 0.04182$; with $R^2 = 0.998$ and a residual sum of squares equal to 0.094.

These three fits are accepted by a Kolmogorov-Smirnov test at the 10% level. See Fig. 3 for a graphical illustration of these excellent results.

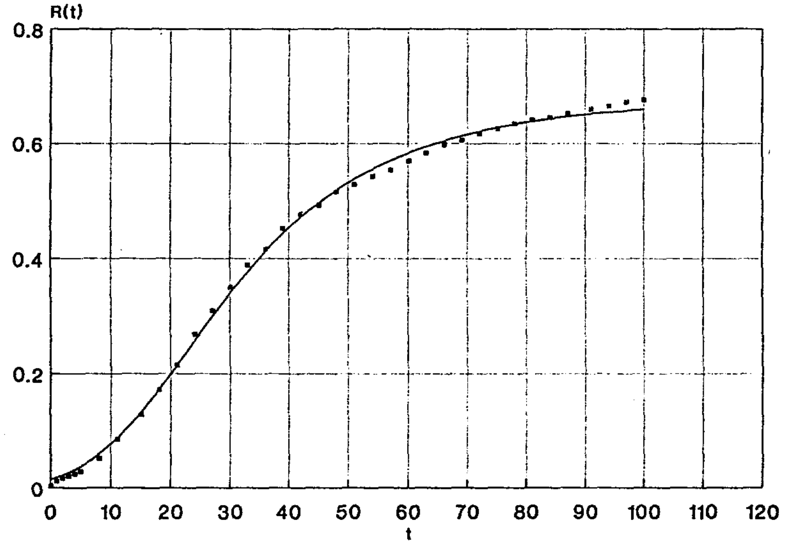

Fig. 3. JACS data. Best fitting curve using the second double exponential model (the same points have been selected as for Fig. 2)

Predictions: 68.4% (leading), 66.5% (communications) and 67.6% (all articles) will eventually be cited. These predictions should be compared with the results in Ref. 19 where it is shown that all articles published in volume 87 (1965) of JACS were eventually cited. Of course, total response, being an extremely rare event, can in this case be explained by the facts that JACS is a very prestigious journal, that citations in all ISI-journals were counted (not only in JACS itself, as we did), and in particular that self-citations were included.

Finally, we check whether this model improves the Motylev-data. For Gompertz' equation (16) applied to the 1960 data, we find:

$M = 0.573$; $b = 0.127$; $a = 0.125$

with $R^2 = 0.967$ and a residual sum of squares equal to 0.0069.

For the 1965 data, we obtain:

$M = 0.546$; $b = 0.0795$; $a = 0.327$

with $R^2 = 0.985$ and a residual sum of squares equal to 0.0048.

Predictions: 36.4% (for the 1960 articles) and 35.3% (for the 1965 articles) will ever be cited by other Russian publications. Both fits are accepted by a Kolmogorov-Smirnov test at the 10% level. Yet, the first model fits slightly better.

Conclusions

We have proposed two double exponential models to describe the first-citation or response distribution and have fitted real data to these models. Moreover, we provided some rationales why these distributions might be appropriate.

For our examples, we have found that when using one year as a time unit, the first model, without inflection, fits the data well. However, when we used two weeks as a unit, we observed an S-shaped curve and needed the second model. Is this a rule or just a coincidence? Or is this difference explained by the fact that for JACS only citations in JACS are counted? This also leads to the problem of how to determine the exact time between publications in different journals. An interesting discussion about this point occurred some years ago in the *Journal of Documentation*. *Windsor*[20] suggested using the publication date of the *Current Contents* issues (a very sensible proposal, I think), but got very strong opposition from *Carmel*.[21] A referee of this article suggested using the date of submission as time zero for determining first citation dates. Note, however, that this date is not always available or clearly defined. It seems that scientists have avoided the use of exact dates, but in the future, more refined models will certainly need more refined data.

*

We would like to thank two anonymous referees for their helpful suggestions. We further thank Lillian *Beltaos* from Mount Saint Vincent University (Halifax, Canada) and Leo *Egghe* (LUC, Diepenbeek, Belgium), for stimulating conversations concerning the research reported in this article. We thank Joke *Margodt* (KIHWV) for making the original figures. Presentation and structure of this article were influenced by the recommendations of *Salager-Meyer*.[22]

References

1. E. GARFIELD, *Citation Indexing – Its Theory and Applications in Science, Technology, and Humanities*, New York, Wiley, 1979.

2. L. EGGHE, R. ROUSSEAU, *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science,* Amsterdam, Elsevier, 1990.
3. B. C. PERITZ, On the objectives of citation analysis: problems of theory and method, *Journal of the American Society for Information Science*, 43 (1992) 448-451.
4. B. C. BROOKES, The foundations of information science. Part IV. Information science: the changing paradigm, *Journal of Information Science*, 3 (1981) 3-12.
5. S. D. NEILL, Brookes, Popper, and objective knowledge, *Journal of Information Science,* 4 (1982) 33-39.
6. E. GARFIELD, Citation behavior – an aid or a hindrance to information retrieval? *Current Contents*, May 1, 1989.
7. D. P. HAMILTON, Publishing by – and for? – the numbers, *Science*, 250, (1990) 1331-1332.
8. G. HERNES, The process of entry into first marriage, *American Sociological Review*, 37 (1972) 173-182.
9. H. F. MOED, A. F. J. VAN RAAN, Cross-field impact and impact delay of physics departments, *Czechoslovak Journal of Physics*, B36 (1986) 97-100.
10. A. SCHUBERT, W. GLÄNZEL, Mean response time – a new indicator of journal citation speed with application to physics journals, *Czechoslovak Journal of Physics*, B36 (1986) 121-125.
11. R. ROUSSEAU, Double exponential models for first-citation processes. Report UIA (1993).
12. G. O. WARE, A general statistical model for estimating future demand levels of database utilization within an information retrieval organization, *Journal of the American Society of Information Science*, 24 (1973) 261-264.
13. V. M. MOTYLEV, 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 (1981) 3-12.
14. D. W. MARQUARDT, An algorithm for least squares estimation of nonlinear parameters, *Journal of the Society of Industrial and Applied Mathematics*, 2 (1963) 431-441.
15. P. F. VERHULST, Recherches mathématiques sur la loi d'accroissement de la population, *Nouveaux Mémoires de l'Académie des Sciences et Belles Lettres de Bruxelles*, 18 (1845) 1-38.
16. A. AKANUMA, Parameter analysis of Gompertz function growth model in clinical tumors, *European Journal of Cancer*, 14 (1978) 681-688.
17. G. F. BRUNTON, T., E. WHELDON, The Gompertz equation and the construction of tumor growth curves, *Cell and Tissue Kinetics*, 13 (1980) 455-460.
18. L. EGGHE, I. K. RAVICHANDRA RAO, Classification of growth models based on growth rates and its applications, *Scientometrics*, 25 (1992) 5-46.
19. J. S. GHOSH, M. L. NEUFIELD, Uncitedness of articles in the Journal of the American Chemical Society, *Information Storage and Retrieval*, 10 (1974) 365-369.
20. D. A. WINDSOR, Assigning dates to articles from dateless journals (letter to the editor), *Journal of Documentation*, 35 (1979) 306.
21. M. J. CARMEL, Dateless journals (letter to the editor), *Journal of Documentation*, 36 (1980) 168-169.
22. F. SALAGER-MEYER, Medical English abstracts: how well are they structured? *Journal of the American Society for Information Science*, 42 (1991) 528-531.