Jointly published by Kluwer Academic Publishers, Dordrecht and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 48, No. 3 (2000) 345–359

A heuristic study of the first-citation distribution

LEO EGGHE[1,2]

[1]*LUC, Diepenbeek (Belgium)*
[2]*UIA, Wilrijk (Belgium)*

The first-citation distribution, i.e. the cumulative distribution of the time period between publication of an article and the time it receives its first citation, has never been modelled by using well-known informetric distributions. An attempt to this is given in this paper. For the diachronous aging distribution we use a simple decreasing exponential model. For the distribution of the total number of received citations we use a classical Lotka function. The combination of these two tools yield new first-citation distributions.

The model is then tested by applying nonlinear regression techniques. The obtained fits are very good and comparable with older experimental results of *Rousseau* and of *Gupta* and *Rousseau*. However our single model is capable of fitting all first-citation graphs, concave as well as S-shaped ; in the older results one needed two different models for it.

Our model is the function

$$\Phi(t_1) = \gamma(1-a^{t_1})^{\alpha-1} \quad .$$

Here γ is the fraction of the papers that eventually get cited, t_1 is the time of the first citation, a is the aging rate and α is Lotka's exponent. The combination of a and α in one formula is, to the best of our knowledge, new. The model hence provides estimates for these two important parameters.

Intorduction

Citation analysis has become a wide-spread discipline, mainly because it is heavily used by science policy and research evaluation professionals. One of the most popular indicators in these studies is the impact factor (IF) derivable from citation analysis, measuring the average number of citations that a journal receives in a two-year period. Another important parameter is the aging rate, a, the decline from year to year of the number of citations a paper receives. For some basic results on IF and a and for their applications we refer the reader to the third part of the book *Egghe* and *Rousseau* (1990).

0138–9130/2000/US $ 15.00

An important indicator of the visibility of research and of the "response times" in a certain discipline is the time t_1 at which an article receives its first-citation. Time t_1 is important for an article since at this time the article shifts its status from "unused" to "used" and the smaller t_1 is, the more we can say – in general – that the article under study is important and early visible in the scientific world. One could also say (cf. *Moed* and *Van Raan* (1986) and *Schubert* and *Glänzel* (1986)) that t_1 is a measure of immediacy, however not directly related to the immediacy index (II), see *Egghe* and *Rousseau* (1990).

Since, in citation analysis, one does not study single articles but homogeneous disciplines one can talk in this connection about "first-citation distributions", i.e. the distribution of the t_1's in the complete discipline. In this paper we will restrict our attention to the cumulative first-citation distribution. It is more common to do so – cf. also *Rousseau* (1994) and *Gupta* and *Rousseau* (1999) – , it is easier (simpler formulae – see further) and, finally, cumulative processes are more important here since their limiting value is the fraction of the eventually cited articles. If we denote by γ this fraction, then $1-\gamma$ is the fraction of the articles that are never cited, an important indicator.

There are not many papers involved in the study of the first-citation distribution. In *Glänzel* (1992) and *Glänzel* and *Schoepflin* (1995) one studies i^{th} Harmonic Mean Response Times, being the harmonic means of the time elapsed between the publication date and the date of the i^{th} (i=1,2,...) citation of the papers. Of course, i=1 represents first-citation. Basically there is only *Rousseau* (1994) who develops a model for the first-citation distribution. His arguments are based on the definition of two differential equations leading to two different models (the one not implied by the other). One model is based on the differential equation

$$R'(t_1) = Be^{-\beta t_1}(1 - R(t_1)) \tag{1}$$

where $R(t_1)$ is the cumulative relative number of cited articles up to time t_1 and B and β are constants, leading to

$$R(t_1) = 1 - kb^{(1-e^{-\beta t_1})} \tag{2}$$

where k and b are other constants.

This model fits well in situations where the first-citation data are concave in t_1 as is, e.g., the case for the data (derived by *Rousseau*) appearing in *Motylev* (1981) on references in the Russian scientific literature to Russian language library science periodicals. However, as is easy to see, model (2) can only handle concave cases since $R''(t_1)<0$ always. There exist, however, cases where the first-citation process is

S-shaped, i.e. starts in a convex way, then proceeds in a concave way, both parts being separated by an inflection point. In fact, *Rousseau* himself points this out by collecting data on first citations of JACS articles in JACS (JACS = *Journal of the American Chemical Society*). This resulted in an S-shaped cloud of points (see Figs 2 and 3 in *Rousseau* (1994) or see further Fig. 3 where the data have been re-used) for which (2) is unsuited. This lead *Rousseau* to develop a second model based on the differential equation.

$$R'(t_1) = Be^{-\beta t_1}(1 - R(t_1))R(t_1) \tag{3}$$

leading to

$$R(t_1) = 1/(1 + 1/MBe^{-\beta t_1}) \ , \tag{4}$$

where M, b and β are constants. This function is capable of fitting the S-shaped cases very well, as can be seen in *Rousseau* (1994).

We want to make the following remarks on these models. The work of *Rousseau* has some explanatory value since it is derived (in a mathematical way) from differential equations (1) and (3). This technique is custom in sciences like physics and chemistry, giving indications about the dynamics of first-citation processes but often these equations are only expressions of widely accepted and experimentally verified properties. In equation (1), the first-citation rate is proportional with the fraction of the (at that time) yet uncited articles, but with proportionality factor decreasing with t_1. In (3), this rate is multiplied by $R(t_1)$ itself, for which a rationale seems to be missing.

The most important drawback of *Rousseau*'s results is that none of the two models is capable of modelling all existing first-citation relations: the first one is needed to model the concave cases while the second one is needed to model the S-shaped cases. This leads to two different rationales (coming from the different differential equations (1) and (3)) for the two types of first-citation relations. Not that this is wrong in itself but in this way we lack the rationale for this different behavior.

Last but not least, one can wonder if the observed first-citation regularities cannot be explained by using elementary informetric tools. After all we are dealing with citation times which are well-described in the literature (see e.g., *Egghe* and *Rousseau* (1990) or *Egghe* and *Rao* (1992)) and with numbers of citations. The simplest model for the former is the aging distribution.

$$c(t) = ba^t \ , \tag{5}$$

where a and b are constants and where $0<a<1$, $t>0$ and c(t) is the density function of citations to an article, t time after its publication. We underline that in *Egghe* and *Rao*

(1992) a rationale has been given for the validity of the lognormal form for c but that (5) is the basic "pure" decay, giving rise to the well-known aging rate a. In this paper we will work with (5) and see how far we can go. In any case (5) is the function to look at in the first place at least from a mathematical point of view. The study of the lognormal case is postponed to another paper.

Now what about the latter one: the number of citations that a paper receives ? This is a typical frequency law as described by Lotka : sources produce items (see *Egghe* and *Rousseau* (1990)): in this case sources are articles and the citations they generate are the items. Whenever we have such a source-item-relation, one of the simplest and explained frequency laws for the fraction of sources with A items is given by

$$\varphi(A) = \frac{D}{A^{\alpha}} \quad , \tag{6}$$

where D and α are constants (α is called the Lotka exponent, $\alpha>1$). Most classically $\alpha\approx 2$ but values above or below 2 are possible. Note that we use (6) here in the case of article citations where inequality in number of citations is very high: many papers are hardly cited and only a few are cited very often. It is well-known that the more unequal a situation is, the larger α must be (cf. *Egghe* and *Rousseau* (1990)) so that $\alpha>2$ will occur more often than $1<\alpha\leq 2$ (although the latter case is not excluded in practise as well as in the theory). In fact, we will comment on this, when studying the concrete examples in a later section. In (6) we will use A as a continuous variable as an approximation of the real-life situation where $A\in\mathbb{N}$. Of course, as a mathematical model, no approximations are introduced.

In the next section we will elaborate the model for the first-citation distribution using only the simple formulae (5) and (6). The third section then establishes the fit for the three cases studied by *Rousseau* (1994) and *Gupta* and *Rousseau* (1999). It turns out that our single (and simple) model is capable of fitting all cases by an appropriate choice of the parameters a and α.

The model

We fix a bibliography being a general set of documents, usually in a homogeneous scientific field. We will use (5) for the aging distribution of the citations to this bibliography and assume that the same function applies for each individual article in the bibliography. Since c(t) is a distribution over continuous time t it is easy to see that $b=-\ln a>0$, since $0<a<1$.

Suppose that an individual article in the bibliography receives A citations in total. Hence, since c(t) is a distribution over continuous time, t,

$$n(t) = -(\ln a)a^t A \tag{7}$$

is the function, describing the number of citations t time after publication, to this article. As explained in the introduction, (6), the distribution of the A-values ($A \geq 1$), taken as a continuous variable, is of Lotka-type: here φ denotes the fraction (of the *cited* articles) with A citations. Since (6) is a distribution it is easy to see that $D=\alpha-1$. Since we will also consider articles without any citation we have that

$$\gamma = \frac{\alpha - 1}{A^{\alpha}} \tag{8}$$

denotes the density (of *all* the articles) with A citations. Here $0<\gamma<1$ is the fraction of ever cited articles ; hence $c_0=1-\gamma$ is the fraction of uncited articles.

We have the following theorem:

Theorem: If we have an exponential aging function as in (7) and a Lotka frequency function as in (8) then we have that

$$\Phi(t_1) = \gamma(1 - a^{t_1})^{\alpha - 1} \tag{9}$$

is the cumulative first-citation distribution of the bibliography.

This function is concave iff $1<\alpha\leq 2$ and is S-shaped iff $\alpha>2$.

Proof: For an article with A citations in total, the expected time t_1 at which it receives its first citation is given by the equation

$$\int_0^{t_1} n(t)dt = 1 \ .$$

This yields (denoting A by A_t to underline the t-dependence)

$$A_t = \frac{1}{1 - a^{t_1}} \ . \tag{10}$$

[Note that it also follows that

$$t_1 = \frac{\ln(1 - \frac{1}{A_t})}{\ln a} \tag{11}$$

but we do not need this here].

The cumulative fraction of *all* articles that are already cited at time t_1 (that is, in this approximation, the fraction of articles cited by time t_1 will be the fraction of articles whose total citation rate exceeds A_t with A_t defined in (10)) is given by

$$\int_{A_t}^{\infty} \gamma \frac{\alpha - 1}{x^{\alpha}} dx \quad , \tag{12}$$

with A_t replaced by (10). Indeed, taking cumulation from $t_1=0$ on, (10) shows that the interval $[0,t_1]$ is bijective with the interval $[A,\infty[$ of number of citations. Now (12) is equal to $\gamma A^{1-\alpha}$ and hence, using (10), we have that the cumulative first-citation distribution is given by

$$\Phi(t_1) = \gamma(1 - a^{t_1})^{\alpha - 1} \quad . \tag{9}$$

It is easy to see that $\Phi'(t_1)>0$ for all t_1 (obviously) and that $\Phi''(t_1)=0$ iff

$$t_1 = \frac{\ln(\frac{1}{\alpha - 1})}{\ln a} \quad . \tag{13}$$

This shows that Φ is entirely concave for $1<\alpha\leq 2$ and that for all $\alpha>2$, $t_1>0$ in (13) showing that the function Φ'' changes sign on the considered $\mathbb{R}^+$ line for t_1. This change of sign must be from positive to negative since $\Phi(0)=0$ and

$$\lim_{t_1 \to \infty} \Phi(t_1) = \gamma < \infty \quad . \tag{14}$$

This proves the S-shape iff $\alpha>2$.

Note: That an S-shaped curve occurs only for large values of α and that the S-shape becomes more and more apparent the larger α is (as is clear from (13) and also from examples we have been drawing) is intuitively clear. The higher α the more inequality we have (in this case between the different total number of citations per article) – see *Egghe* and *Rousseau* (1990), so, relatively speaking, there are fewer articles with a large number A of citations. In other words, using (11), there are not many cases of low values of t_1 and hence the first-citation process starts in a convex way, meaning that Φ increases very slowly.

This note makes clear the involvement of Lotka's α in the study of the first-citation process, besides the aging rate a, whose involvement in the first-citation process is much more evident.

Function (9) is remarkably simple and it still needs to be seen if it is capable of fitting practical situations. This is done in the next section. Note that, once this is done, we receive an estimate of a and α, two of the most important parameters in informetrics. The number a is the basis for all aging studies and α is the basis for concentration (inequality) and other informetrics studies of the bibliography.

Fitting first-citation data

We re-use the data collected in *Rousseau* (1994) and *Gupta* and *Rousseau* (1999), with permission. We start with the latter one.

Gupta and Rousseau data

Source articles were taken from the "Bibliography of Theoretical Population Genetics". Nine databases were constructed all with similar concave first-citation shapes. The one depicted in *Gupta* and *Rousseau* (1999) is the set of articles published in 1973 (418 source articles). That is why we will also use it here. The data are given in Table 1.

Table 1
First-citation data of 1973 articles

Year	Number of articles cited for the first time	Fraction of column 2	Cumulative fraction of first-citation
1973	38	0.0916	0.0916
1974	72	0.1736	0.2652
1975	69	0.1664	0.4316
1976	25	0.0603	0.4919
1977	21	0.0506	0.5425
1978	16	0.0386	0.5811
1979	8	0.0193	0.6004
1980	9	0.0217	0.6221
1981	4	0.0096	0.6317
1982	6	0.0145	0.6462
1983	4	0.0096	0.6558
1984	5	0.0121	0.6679
1985	1	0.0024	0.6703
1986	2	0.0048	0.6751
1987	2	0.0048	0.6799

Note that about 68% of the articles is cited at least once. We have fitted these data in two ways. Once we used

$$\Phi(t_1) = \gamma(1 - a^{t_1})^{\alpha - 1} \tag{9}$$

as a 2-parameter distribution, putting γ=0.68. Once we used (9) as a 3-parameter distribution. Each time the parameters that must be calculated by the system (STATGRAPHICS 7.1) have to be estimated. This was not always easy. To get a feeling of the magnitude of the estimates it is advisable to draw some graphs (9) for different parameters. We executed this in the program MATHCAD 4.0.

With the 2-parameter fitting we found (putting γ=0.68) a=0.672 and α=2.536. With the 3-parameter fitting we found a=0.635, α=2.756 and γ=0.665. Note that this last value is less than 0.6799, the highest value in the fourth column of table 1. This shows that the non-linear regression fitting is not capable of calculating the real prospect of $\Phi(t_1)$ for t_1 going to ∞. It is a fitting device and since in both cases R^2 is over 0.99 and since the visual inspection of the 2-parameter fitting is at least as good as the one of the 3-parameter fitting, we keep the former one. Note that – from an explanatory point of view – it is best to have models with the least number of parameters. The result of the 2-parameter fitting is shown in Fig. 1.

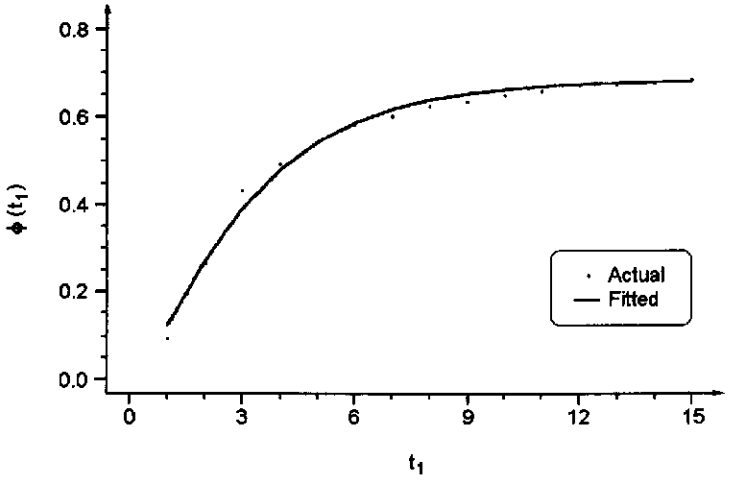

Fig. 1.Two-parameter fitting of (10) to the *Gupta* and *Rousseau* data

The fitting is of the same high quality as in *Gupta* and *Rousseau* (1999) but using only 2 parameters a and α [of course we are convinced that, although the *Gupta* and *Rousseau* model has 3 parameters, they would also be able to fit well with 2 parameters, by estimating their equivalent of γ].

Motylev data (appearing in Rousseau (1994))

These citation data concern references in the Russian scientific literature to Russian language library science periodicals, published by *Motylev* (1981). The data were re-used in *Rousseau* (1994). We will use them here for testing our model. The data are presented in Table 2.

Table 2
First-citation data of Motylev

Year	Number of articles cited for the first time	Fraction of column 2	Cumulative fraction of first-citation
1	10	0.018	0.018
2	33	0.060	0.078
3	18	0.033	0.111
4	17	0.031	0.142
5	14	0.026	0.168
6	4	0.007	0.175
7	0	0.000	0.175
8	10	0.018	0.193
9	7	0.013	0.206
10	7	0.013	0.219
11	8	0.015	0.234
12	9	0.016	0.250
13	4	0.007	0.257
14	13	0.024	0.281
15	6	0.011	0.292
16	6	0.011	0.303

Note that in this case only about 30% of all articles is cited. Here a substantial improvement occured when using a 3-parameter fit above a 2-parameter fit. We obtained for (9) a=0.956, α=1.746, γ=0.486. The latter value is far above 0.303 which is clear by visual inspection of the graph: after 16 years the graph is far from being horizontal (contrary to e.g., Fig. 1); hence adding further years will result in more cited papers (increasing the low number of 30%!). The calculated model is shown in Fig. 2. The fit is very good, giving an R^2 of over 0.97.

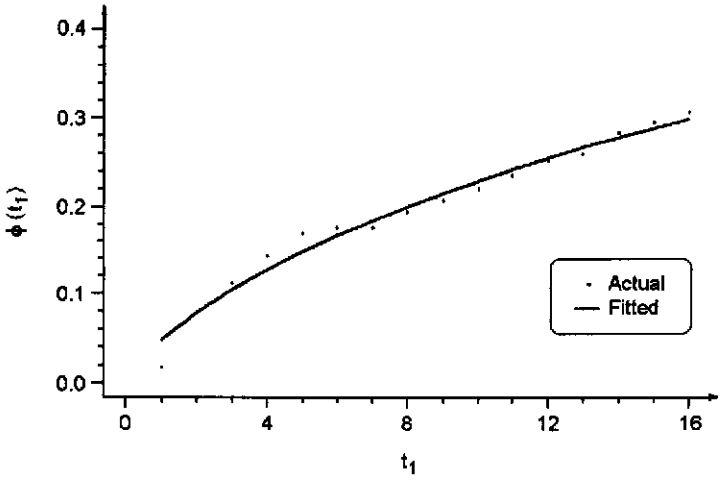

Fig. 2. Three-parameter fitting of (9) to the *Motylev* data

JACS to JACS data of Rousseau (1994)

These data were collected by *Rousseau* in JACS of the year 1975. The first-citatons to these articles are followed in JACS during a 4-year period, hereby checking 102 issues since JACS publishes issues on a biweekly basis. We present the data in Table 3 in the Appendix. Note that, after 4 years, 67.6% of the articles is cited.

We have checked the 2-and 3-parameter fits and found only small differences. We, therefore, present the 2-parameter model (for γ=0.676): a=0.955, α=3.641. We have a very good fit: R^2 is 0.999 ! See Fig. 3 for graphical inspection.

The quality of the fit is very good and about the same as the one found in *Rousseau* (1994) (but there using another mathematical model than in the cases of Fig. 1 and 2).

Note: In addition to first-citation distributions we also obtained values for Lotka's α. We see that $\alpha<2$ in case of the *Motylev* data and that $\alpha>2$ in case of the *Gupta-Rousseau* and *Rousseau* data. It is not surprising that these values are found: the higher α, the more concentrated (i.e., unequal) the distribution is and it is well-known that this occurs more in the sciences (e.g., genetics, chemistry) than in other disciplines (e.g., library science). The found α-values are in accordance to this (cf. *Egghe* and *Rousseau* (1990)).

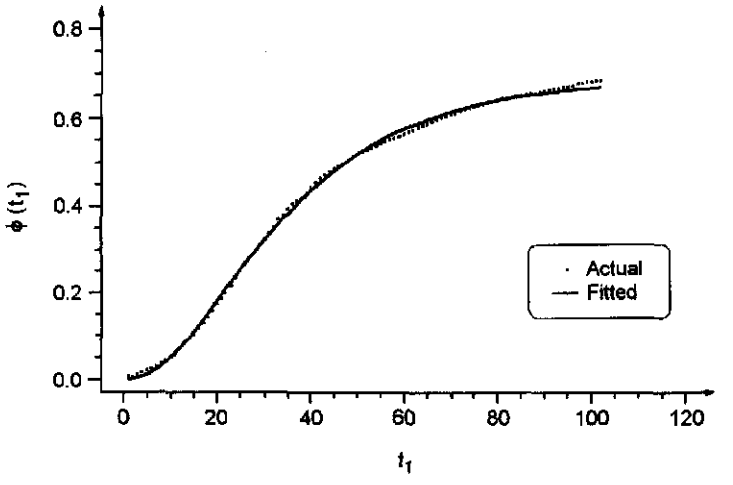

Fig. 3. Two-parameter fitting of (10) to the *Rousseau* data

Conclusions and further research

By combining an exponentially decreasing aging function for the citations to articles and a Lotka function for the number of received citations, we were able to prove the following mathematical model for the cumulative distribution of first-citation times:

$$\Phi(t_1) = \gamma(1-a^{t_1})^{\alpha-1} \tag{9}$$

combining the aging rate a and Lotka's exponent α in one formula. This yields an informetric rationale for the first-citation distribution.

We have shown by practical examples that function (9) on its own is capable of fitting accurately first-citation data, whether they are concavely shaped or S-shaped. *Rousseau* (1994) needed two different models for this. This fact and the fact that the model only uses elementary informetric tools (aging rate a and Lotka's exponent α) makes us conclude that model (9) is to be preferred above *Rousseau*'s models.

It is remarkable how well (9) can fit data, although we used an exponentially decreasing aging model. It is well-known that real aging curves are modelled e.g., by a lognormal distribution – see *Matricciani* (1991) and *Egghe* and *Rao* (1992). In a forthcoming paper we hope to even refine the above model by combining the lognormal aging distribution with the Lotka production distribution.

*

The author is grateful to prof. Dr. R. *Rousseau* for interesting discussions on the topic of this paper.

References

EGGHE, L., I. K. RAVICHANDRA RAO (1992), Citation age data and the obsolescence function: fits and explanations. *Information Processing and Management* 28(2): 201-217.

EGGHE, L., R. ROUSSEAU (1990), *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam.

GLÄNZEL, W. (1992), On some stopping times of citation processes. From theory to indicators. *Information Processing and Management*, 28: 53-60.

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

GUPTA, B.M., R. ROUSSEAU (1999), Further investigations into the first-citation process: the case of population genetics. Libres, 9(2), aztec.lib.utk.edu/libres/libre9n2/fc.htm.

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

MOED, H. F., A. F. J. VAN RAAN (1986), Cross-field impact and impact delay of physics departments. *Czechoslovak Journal of Physics*, B36: 97-100.

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.

ROUSSEAU, R. (1994), Double exponential models for first-citation processes. *Scientometrics*, 30: 213-227.

SCHUBERT, A., W. GLÄNZEL (1986), Mean response time – a new indicator of journal citation speed with application to physics journals. *Czechoslovak Journal of Physics*, B36: 121-125.

Appendix

Table 3
First-citation data of *Rousseau*

Issue	Number of articles cited for the first time	Fraction of column 2	Cumulative fraction of first-citation
1	1	0.0006	0.0006
2	6	0.0033	0.0039
3	14	0.0078	0.0117
4	8	0.0044	0.0161
5	8	0.0044	0.0205
6	7	0.0039	0.0244
7	8	0.0044	0.0288
8	14	0.0078	0.0366
9	10	0.0055	0.0421
10	16	0.0089	0.0510
11	10	0.0055	0.0565
12	26	0.0144	0.0709
13	25	0.0139	0.0848
14	16	0.0089	0.0937
15	16	0.0089	0.1026
16	26	0.0144	0.1170
17	21	0.0116	0.1286
18	26	0.0144	0.1430
19	25	0.0139	0.1569
20	27	0.0150	0.1719
21	23	0.0128	0.1847
22	23	0.0128	0.1975
23	30	0.0166	0.2141
24	31	0.0172	0.2313
25	32	0.0177	0.2490
26	33	0.0183	0.2673
27	22	0.0122	0.2795
28	20	0.0111	0.2906
29	32	0.0177	0.3083
30	23	0.0128	0.3211
31	28	0.0155	0.3366
32	22	0.0122	0.3488
33	34	0.0189	0.3677
34	14	0.0078	0.3755
35	24	0.0133	0.3888
36	23	0.0128	0.4016

(Continued on next page)

(Continued from previous page)

Issue	Number of articles cited for the first time	Fraction of column 2	Cumulative fraction of first-citation
37	11	0.0061	0.4077
38	14	0.0078	0.4155
39	21	0.0116	0.4271
40	24	0.0133	0.4404
41	21	0.0116	0.4520
42	16	0.0089	0.4609
43	14	0.0078	0.4687
44	14	0.0078	0.4765
45	11	0.0061	0.4826
46	11	0.0061	0.4887
47	7	0.0039	0.4926
48	13	0.0072	0.4998
49	14	0.0078	0.5076
50	14	0.0078	0.5154
51	7	0.0039	0.5193
52	6	0.0033	0.5226
53	10	0.0055	0.5281
54	8	0.0044	0.5325
55	9	0.0050	0.5375
56	8	0.0044	0.5419
57	10	0.0055	0.5474
58	4	0.0022	0.5496
59	7	0.0039	0.5535
60	9	0.0050	0.5585
61	10	0.0055	0.5640
62	10	0.0055	0.5695
63	13	0.0072	0.5767
64	3	0.0017	0.5784
65	8	0.0044	0.5828
66	9	0.0050	0.5878
67	8	0.0044	0.5922
68	9	0.0050	0.5972
69	4	0.0022	0.5994
70	6	0.0033	0.6027
71	5	0.0028	0.6055

(Continued on next page)

(Continued from previous page)

Issue	Number of articles cited for the first time	Fraction of column 2	Cumulative fraction of first-citation
72	10	0.0055	0.6110
73	5	0.0028	0.6138
74	5	0.0028	0.6166
75	3	0.0017	0.6183
76	6	0.0033	0.6216
77	6	0.0033	0.6249
78	7	0.0039	0.6288
79	5	0.0028	0.6316
80	5	0.0028	0.6344
81	6	0.0033	0.6377
82	3	0.0017	0.6394
83	2	0.0011	0.6405
84	4	0.0022	0.6427
85	1	0.0006	0.6433
86	3	0.0017	0.6450
87	4	0.0022	0.6472
88	1	0.0006	0.6478
89	6	0.0033	0.6511
90	4	0.0022	0.6533
91	5	0.0028	0.6561
92	3	0.0017	0.6578
93	2	0.0011	0.6589
94	4	0.0022	0.6611
95	2	0.0011	0.6622
96	5	0.0028	0.6650
97	6	0.0033	0.6683
98	3	0.0017	0.6700
99	4	0.0022	0.6722
100	1	0.0006	0.6728
101	4	0.0022	0.6750
102	2	0.0011	0.6761

Received March 27, 2000.

Address for correspondence:
LEO EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek (Belgium)
E-mail: leo.egghe@luc.ac.be