Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 71, No. 2 (2007) 339–348
DOI: 10.1007/s11192-007-1662-7

Parameter identification of the observed citation distribution

GUANG YU, YI-JUN LI

School of Management, Harbin Institute of Technology, Harbin (P. R. China)

Based on the transfer function model of the observed citation distribution and the expression of the cumulative citation probability distribution, parameters of 12 citation distributions are identified from statistical data of age distributions of references of 10 journals in JCR using the parameter optimization fitting method. At same time, based on the steady state solution of differential equations of the publication delay process and data of publication delays of 10 journals, the publication delay parameters of every journal are identified using the fitting method. Identified parameters of every journal citation distribution are compared with the journal's publication delay parameters and some valuable conclusions are deduced.

Introduction

LUWEL & MOED (1997) defined the publication delay as the time between the submission of a manuscript and the actual publication. They studied the influence of publication delays on the aging of scientific literature, and proposed that the publication delay of a journal from the age distribution of its references may possibly be estimated (LUWEL & MOED, 1997). According to the research-citation cycle pictured by EGGHE & ROUSSEAU (2000, p. 159), it is obvious that the age distribution of references of a journal is influenced by its publication delays. Based on the convolution formula of the

Received September 11, 2006

Address for correspondence:
YU, GUANG
Library, Harbin Institute of Technology
No.17, Silin Street, NanGang, Harbin, 150001, Heilongjiang, P. R. China
E-mail: yug@hit.edu.cn

disturbed aging distribution (EGGHE & ROUSSEAU, 2000) and the steady state solution of the mathematic model of publication delay process (YU et al., 2000, pp. 411–412), YU et al. (2005) established the transfer function model of the disturbed citing process by the system identification method and proved the inverse relationship between the field (or discipline) average publication delay and the journal impact factor in theory. In this paper, we will use the model and the analytical expression of the cumulative citation probability distribution to identify three model parameters from data of citation distributions of journals in JCR, respectively, and then we compare parameters identified from the citation distribution with actual publication delay parameters of the journals and hope to obtain some valuable results.

The parameter identification method of the citation distribution

Based on the study of the influence of publication delay on the observed aging distribution of scientific literature by EGGHE & ROUSSEAU (2000), YU et al. (2005) proposed the notion of *the delay effect of literature citation*, namely an actual citing behavior occurs at the same time of creating manuscripts not publishing and the delay effect of literature citation undoubtedly is transferred to the disturbed citation distribution and can be reflected by the disturbed citation distribution, and then established the transfer function model of the disturbed citing process. It is

$$W(s) = \frac{e^{-\tau s}}{(T_1 s + 1)(T_s s + 1)} \quad . \tag{1}$$

In Eq. (1), s is the Laplace variable; T_1 is called as time constant; T_s is another time constant related with the average publication delay, τ is the pure publication delay. In that paper (YU et al., 2006), according to inverse Laplace transform, we deduced the analytical expression of the citation distribution function and then deduced the cumulative citation distribution function $C(T)$ by the integral of the citation distribution function:

$$C(T) = 1 - \frac{T_1 \cdot e^{-\frac{T-\tau}{T_1}}}{T_1 - T_s} + \frac{T_s \cdot e^{-\frac{T-\tau}{T_s}}}{T_1 - T_s} \quad . \tag{2}$$

According to the convolution formula of the disturbed citation distribution (EGGHE & ROUSSEAU, 2000), Eq. (2) describes the relation of the age distribution of cumulative citation probability with the time constant T_1 of the ageing process, the time constant T_s of the publishing process and the pure delay τ; and it provides an effective mathematical tool for the quantificationally study on the actual citation distribution and its impact factors. In Eq. (2), T is the citation age, T_1 is the time constant related with the ageing process and has a relationship with the ageing coefficient α ($T_1 = 1/\alpha$); T_s is another time constant related with the average publication delay; τ is the pure delay. The journal average publication delay indicator is $\overline{T} = \tau + T_s$, $T_s = N/Y$ and N is the deposited contribution quantity of the journal at steady state, Y is the published contribution flux at steady state (YU et al., 2004). The bigger T_s or τ of a journal, the more serious the deposited contribution quantity or the publication delay of the journal, the worse influence on the citation distribution (YU et al., 2006).

In this paper, $C(T)$ is regarded as the analytical value of the cumulative citation distribution probability function (dimensionless) and should be fitted with the statistic data of cumulative% of the journal citations from JCR; $\hat{C}(T_k)$ is regarded as statistical data of the cumulative citation distribution probability. The parameter optimization fitting method is used to identify optimum model parameters (LJUNG, 1999, p. 501). It is an optimization problem to estimate T_1, T_s and τ, the aim is to minimize the Objective Function:

$$J = \sum_{k=1}^{m} (\hat{C}(T_k) - C(T_k))^2 \ . \tag{3}$$

In Eq. (3), m is the quantity of sample points. For validating the model, we choose the classical age distribution data of 80,005 citations of 2,595 papers (published in 1980) from *Journal of Biological Chemistry* and *Biochemistry* (the total of citations is 85,431 items, journal papers account for 93.6%) (WANG, 1997) as statistical data $\hat{C}(T_k)$ (see Table 1). Let k as sample points: $k = 1$ to 30, $m = 30$, the age T_k varies from 1 to 30 years. Based on Eq. (2), model parameters of the citation distribution are identified. In Figure 1, the dot curve is the statistic data, the solid curve is the simulating result. We can see that this model is fit for the citing process and the simulating result is satisfactory. According to analyses of the aging process and publishing process, bigger one of two time constants identified from the citation distribution data is the time constant related with the aging process and another is the time constant related with the publishing process. Three identified parameters are $T_1 = 6.2430$years, $T_s = 0.4323$years, $\tau = 0.6525$years, respectively. According to our thought in modeling, T_s and τ reflect the average publication delay of 2,595 papers of the two journals published in 1980: $\overline{T} = T_s + \tau = 1.0848\text{years}$.

Finally we should validate the fitting accuracy of the model from the fitting error $\Delta C(T)$,

$$\Delta C(T) = C(T) - \hat{C}(T) \ . \tag{4}$$

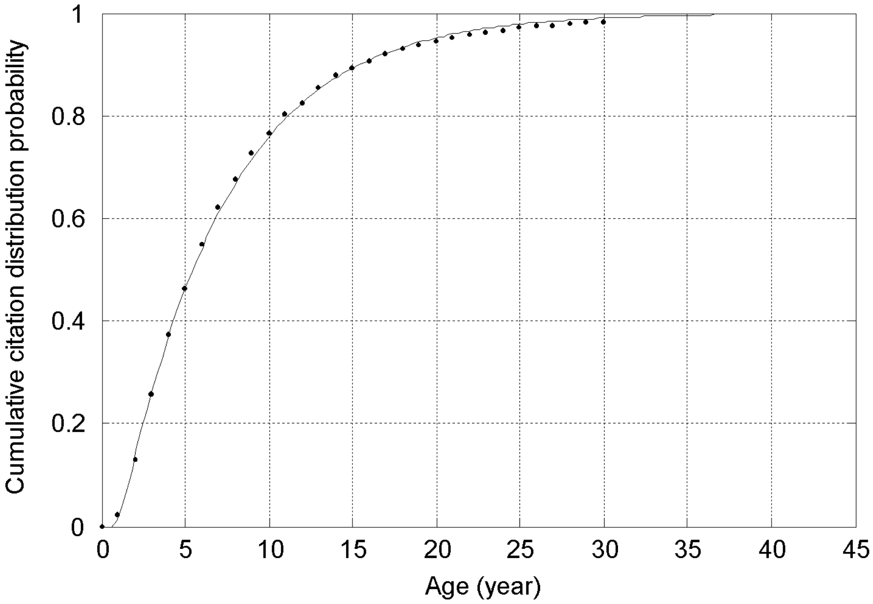

Figure 1. The age distribution probability curve of cumulative citations of biochemistry documents

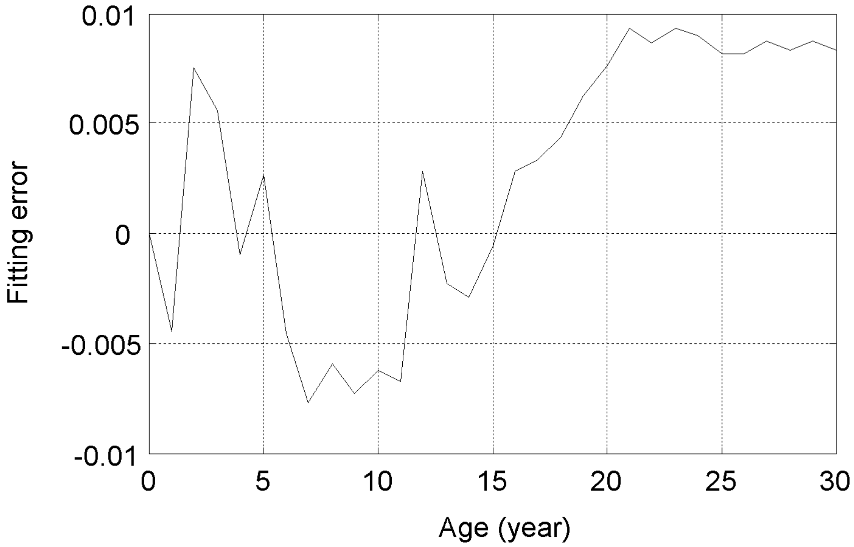

Figure 2. The fitting error curve

Figure 2 shows the fitting error distribution curve. We can see that the maximal error is not over 1.2% and the model can be used to analyze theoretically the citation distribution law of a journal.

Because we are incapable to find the two journals in our library, data of their publication delays cannot be obtained; so it cannot be validated whether the time constant T_s and the pure delay τ identified from the disturbed citation distribution are coincident with parameters of actual publication delays. In this paper, we choose stochastically 10 journals and collect publication delay data of these journals; then based on the steady state solution of the publishing process (YU et al., 2000, pp. 411–412), parameters (T_s and τ) of every journal's publishing process are solved and compared with three parameters (T_1, T_s and τ) of every journal identified from its citation distribution data in JCR using the parameter optimization fitting method given above.

Table 1. Citation distribution data of two biochemistry journal in 1980

year	T_k	Cumulative citations quantity	$\hat{C}(T_k)$	year	T_k	Cumulative citations quantity	$\hat{C}(T_k)$
1980	1		0.0236	1963	18	74330	0.9291
1979	2	10413	0.1302	1962	19	74927	0.9371
1978	3	20535	0.2567	1961	20	75530	0.9441
1977	4	29780	0.3722	1960	21	75969	0.9496
1976	5	37331	0.4617	1959	22	76512	0.9563
1975	6	43866	0.5483	1958	23	78675	0.9609
1974	7	49535	0.6191	1957	24	77249	0.9656
1973	8	53997	0.6749	1956	25	77624	0.9702
1972	9	58039	0.7254	1955	26	77876	0.9734
1971	10	61293	0.7661	1954	27	78053	0.9756
1970	11	64180	0.8022	1953	28	78265	0.9783
1969	12	66569	0.8230	1952	29	78399	0.9799
1968	13	68317	0.8539	1951	30	78566	0.9820
1967	14	70129	0.8765	1950-1941	31-40	79255	0.9906
1966	15	71438	0.8929	1940-1921	41-60	79910	0.9988
1965	16	72445	0.9055	Before 1920	Over 61	80005	1.000
1964	17	73493	0.9186				

Choosing journals and collecting their publication delay data

In the paper the publication delay is defined as time interval between submission and publication of accepted papers**.** We choose stochastically 10 journals and collect data of their publication delays, volumes and issues of these journals and numbers of statistical papers are:

1) *International Journal of Engineering Science* (shortened form: IJES) 2003, (1-12), 120 items;
International Journal of Engineering Science (shortened form: IJES) 2000, (1-18), 111 items;
2) *Journal of Mathematical Physics* (shortened form: JMP) 2001, (1-12), 235 items;
3) *Mechanics Research Communications* (shortened form: MRC) 2000, (1-4), 95 items;
4) *Journal of the European Ceramic Society* (shortened form: JECS) 2002, (1-12), 269 items;
5) *Scientometrics* (shortened form: Sciento) 2002, 61items;
6) *IEEE Transactions on Automatic Control* (shortened form: IEAC), 2003, (1-5), 146 items;
7) *Information Process & Management* (shortened form: IPM), 2002, (1-6), 91 items;
8) *Journal of Applied Physics* (shortened form: JAP), 2001, 89 (1-12), 706 items;
Journal of Applied Physics (shortened form: JAP), 2002, 91(1-12), 1113 items;
9) *Computer-aided Design* (shortened form: CAD), 2002, (1-12), 65 items;
10) *Computers & Education* (shortened form: C&E), 2002, (4-12), 36 items.

Based on the model of the journal publishing process, YU et al. (2000, pp. 411–412) deduced the steady state solution of the process under the reasonable hypotheses: the delay (or age) distribution function of the journal accumulated publication flux, under the particular condition – the publication probability density is a constant, namely every contribution's publication probability is same. This solution is

$$\int_0^T y(T)dT = Y(1\text{-}e^{-\frac{(T-\tau)}{T_s}}) \ . \tag{5}$$

In Eq. (5), $T_S = N/Y$, Y is the published contribution flux at steady state (the quantity of published contributions of a journal in one year); then let Eq. (5) is divided by Y and

$$Y(T) = \frac{1}{Y}\int_0^T y(T)dT \ ,$$

we obtained the expression of the cumulative published probability function $Y(T)$:

$$Y(T) = 1 - e^{-\frac{(T-\tau)}{T_s}} \ . \tag{6}$$

In Eq. (6), T is the literature age (or delay), T_S is the time constant of the publishing process, τ is called as the pure delay. The parameter optimization fitting method is used yet to solve the two parameters from publication delay data of a journal too. Using the expression: $\overline{T} = \tau + T_S$, the average publication delay of every journal can be accounted (see Table 2). Dimensions of $\overline{T}$, T_S and τ are *year*.

Collecting data of citation distributions and identifying model parameters

We collect data of citation distributions of 10 journals from JCR of ISI. 12 groups of data are listed in the Appendix in the end of this paper. Every data table is breakdown of the citations from every journal by the cumulative percent of statistical year are as follows:

1) IJES: Citation distribution data of 2003: Citing Half-Life: >10.0 years
Citation distribution data of 2000: Citing Half-Life: >10.0 years
2) JMP: Citation distribution data of 2001: Citing Half-Life: >10.0 years
3) MRC: Citation distribution data of 2000: Citing Half-Life: >10.0 years
4) JECS: Citation distribution data of 2002: Citing Half-Life: 9.2 years
5) Sciento: Citation distribution data of 2002: Citing Half-Life: 8.1 years
6) IEAC 2003: Citation distribution data of 2003: Citing Half-Life: 7.8 years
7) IPM: Citation distribution data of 2002: Citing Half-Life: 7.6 years
8) JAP: Citation distribution data of 2002: Citing Half-Life is 6.7 years
Citation distribution data of 2001: Citing Half-Life is 6.7 years
9) C & E: Citation distribution data of 2002: Citing Half-Life: 6.6 years
10) CAD: Citation distribution data of 2002: Citing Half-Life is 8.1 years

Based on the parameter optimization fitting method of the cumulative citation distribution introduced above, 12 cumulative citation distributions of 10 journals are identified and three parameters of every distribution are solved, then every average publication delay $\overline{T} = \tau + T_S$ is accounted, see Table 2. But citation distribution data of 10 years is only listed in JCR; the fitting error of the model is bigger than the citation distribution of two biochemistry journals. For improving the fitting precision, we revise identified model parameters using the model (Eq. (2)) to identify parameters from data of the age distribution of references of every journal (second row in each data table). Because this calculating process is complex and difficult for many readers to understand, we think there is no need to describe the process in this paper. Two groups of identified parameters are shown in Table 2 and some valuable results are as follows:

1. Two identified time constant (T_1 and T_S) are always big discrepant on their size, the bigger one is regarded as the time constant related with the aging process and the smaller one as the time constant related with the publishing process.
2. The longer the journal Citing Half-Life, the bigger the time constant T_1 identified from the observed citation distribution data.

3. The time constant T_s identified from the observed citation distribution data of a journal is different from another T_s of the journal's actual publication delay, see Table 2; and that those journals with smaller Citing Half-Life, there is much difference between two T_ss, such as those journals (JAP, C&E, CAD) in Table 2.
4. The time constant T_1 identified from citation distribution of same journal in different year has no (or slight) change. It is because this ageing time constant is related with the ageing coefficient of the academic subject that the journal belongs to.
5. Parameters T_s and τ identified from citation distribution are influenced by the publication delay of the journal, namely the identified average delay $T = T_s + \tau$ from the citation data would prolong too when the journal publication delay increases; in Table 2, $\overline{T}$ of IJES (2003) & IJES (2000) and JAP (2001) & JAP (2002) show this characteristic.

Table 2. Two group of identified parameters

Rank	Journal title abbreviation (year)	Citing half-life	Identified parameters of the citation distribution data					Identified parameters of the publishing process		
			T_1	T_s	τ	$\overline{T}$	ΔC_{max}	T_s	τ	$\overline{T}$
1	IJES (2003)	>10	17.165	0.644	0.447	1.091	0.0055	0.635	0.308	0.943
	IJES (2000)	>10	16.205	1.217	1.015	2.232	0.0085	0.963	0.575	1.538
2	JMP (2001)	>10	14.919	0.618	0.236	0.854	0.0134	0.418	0.397	0.815
3	MRC(2000)	>10	13.224	0.509	0.303	0.812	0.0092	0.552	0.212	0.764
4	JECS (2002)	9.2	10.458	0.998	0.934	1.932	0.0052	1.126	0.253	1.379
5	Sciento (2002)	8.1	9.496	0.632	0.483	1.115	0.0220	0.298	0.213	0.511
6	IEAC (2003)	7.8	8.702	1.290	0.297	1.587	0.0119	0.724	0.544	1.268
7	IPM (2002)	7.6	7.992	1.365	0.652	2.017	0.0120	0.877	0.497	1.374
8	JAP (2001)	6.7	7.903	0.630	0.539	1.169	0.0152	0.409	0.258	0.667
	JAP (2002)	6.7	7.801	0.603	0.612	1.215	0.0131	0.401	0.264	0.665
9	C&E(2002)	6.6	7.412	0.902	0.655	1.557	0.0221	0.813	0.237	1.050
10	CAD(2002)	8.1	6.481	1.856	1.523	3.379	0.0069	1.373	0.431	1.804

Conclusions

According to two groups of identified parameters in Table 2, we can get some conclusions:

1. According to analyzing the ageing process and the publishing process, bigger one of two time constants (T_1 and T_s) identified from citation distribution data is the time constant T_1 related with the citing process and smaller one is the time constant T_s related with the journal publication delay.

2. The average publication delay ($\overline{T} = T_S + \tau$) identified from citation distribution data of a journal is longer than the journal's average publication delay in statistical year. According to the research-citation cycle (EGGHE & ROUSSEAU, 2000, Figure 1), the parameters (T_S and τ) of the citation distribution are influenced by some factors, such as the journal's publication delays, the citing behavior mode of authors, the transmitting process of the journal, errors created in the identifying process; however the publication delay is a main factor.
3. The delay parameters (T_S and τ) identified from citation distribution data of the same journal vary with the actual publication delay of the journal; the longer the publication delay is, the bigger the delay parameters (T_S and τ) are.
4. According to the observed citation distribution model (Eq. (1)) and the identified results of journal citation distributions, besides the ageing factor of the academic subject that the journal belong to, main factors which affect the citation distribution of a journal include publication delays of the journal in statistical year, too; but the identified average delay ($\overline{T} = T_S + \tau$) is longer than the journal's average publication delay, it is showed that the identified parameters (T_S and τ) from the citation data also reflect effect of other impact factors (such as the transmitting process of literature) besides the journal publication delay; this result theoretically negates the idea of LUWEL & MOED (1997) that publication delays of a journal can be estimated from the age distribution of its citations.

*

Authors would like to thank Professor Da-Ren Yu for his comment and assistance on data processing.

This work is supported by grant G70571020 from the National Nature Science Foundation of China and the TPM center of Harbin Institute of Technology.

References

EGGHE, L., ROUSSEAU, R. (2000), The influence of publication delays on the observed aging distribution of scientific literature, *Journal of the American Society for Information Science,* 51 (2) : 158–165

LJUNG, L. (1999), *System Identification –Theory for the User* (Second edition), Prentice Hall PTR, New Jersey.

LUWEL, M., MOED, H. F. (1998), Publication delays in the science field and their relationship to the aging of scientific literature, *Scientometrics*, 41 (1-2) : 26–40.

YU, G., YU, D. R., RONG, Y. H. (2000), The mathematical models of the periodical literature publishing process, *Information Processing & Management*, 36 (3) : 401–414.

YU, G., YU, D. R., LI, Y. J. (2004), The universal equations of periodical average publication delay at steady state, *Scientometrics,* 60 (2) : 121–129.

YU, G., WANG, X. H., YU, D. R. (2005), The influence of publication delays on impact factors in the scientific field, *Scientometrics*, 64 (2) : 235–246.

YU, G., GUO, R., LI, Y. J. (2006), The influence of publication delays on three ISI indicators, *Scientometrics*, 69 (3) : 511–527.

WANG, C. D. (1997), *Introduction of Bibliometrics*, China: Guangxi teachers college publishing company.

Appendix 1
Data table of breakdown of the citations *from every journal* by the cumulative percent of statistical year [from JCR Science Edition]

IJES (2003)	Cited Year	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993-all
	Cites from 2003	117	375	553	528	557	397	347	418	411	290	2457
	Cumulative %	1.10	5.31	11.10	15.60	20.18	24.29	29.05	33.08	36.74	40.99	100
IJES (2000)	Cited Year	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990-all
	Cites from 2000	6	36	74	110	139	92	92	85	101	73	1315
	Cumulative %	0.28	1.98	5.46	10.65	17.19	21.53	25.86	29.86	34.62	38.06	100
JMP (2001)	Cited Year	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991-all
	Cites from 2001	107	556	647	587	481	396	363	352	288	286	5188
	Cumulative %	1.16	7.17	14.16	20.51	25.71	29.99	33.91	37.71	40.83	43.92	100
MRC (2000)	Cited Year	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990-all
	Cites from 2000	19	71	81	63	65	54	42	41	44	45	540
	Cumulative %	1.78	8.45	16.06	21.97	28.08	33.15	37.09	40.94	45.07	49.30	100
JECS (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	23	138	179	178	138	102	96	68	80	55	800
	Cumulative %	1.24	8.67	18.31	27.89	35.33	40.82	45.99	49.65	53.96	56.92	100
Sciento (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	42	236	553	631	541	561	441	467	429	371	3707
	Cumulative %	0.53	3.48	10.41	18.32	25.10	32.13	37.66	43.51	48.89	53.54	100
IEAC (2003)	Cited Year	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993-all
	Cites from 2003	117	375	553	528	557	397	347	418	411	290	2457
	Cumulative %	1.81	7.63	16.20	24.39	33.02	39.18	44.56	51.04	57.41	61.91	100
IPM (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	3	36	104	137	126	99	86	58	61	46	500
	Cumulative %	0.24	3.11	11.39	22.29	32.32	40.21	47.05	51.67	56.53	60.19	100
JAP (2001)	Cited Year	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991-all
	Cites from 2001	811	5124	6565	5740	4990	4391	3521	2947	2620	2203	21262
	Cumulative %	1.35	9.86	20.77	30.31	38.60	45.90	51.75	56.65	61.00	64.67	100
JAP (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	1052	5567	7550	6821	5697	4836	4086	3462	2811	2476	24253
	Cumulative %	1.53	9.65	20.65	30.59	38.90	45.94	51.90	56.95	61.04	64.65	100
CAD (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	5	22	81	162	163	177	134	137	132	111	667
	Cumulative %	0.28	1.51	6.03	15.08	24.18	34.06	41.54	49.19	56.56	62.76	100
C&E (2002)	Cited Year	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992-all
	Cites from 2002	11	86	108	169	136	101	76	67	72	59	435
	Cumulative %	0.83	7.35	15.53	28.33	38.64	46.29	52.05	57.12	62.58	67.05	100