Information Processing & Management Vol. 28, No. 1, pp. 53-60, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

ON SOME STOPPING TIMES OF CITATION PROCESSES. FROM THEORY TO INDICATORS

WOLFGANG GLÄNZEL
Universität Konstanz, Sozialwissenschaftliche Fakultät, Fachgruppe Psychologie,
Pf. 5560, D-7750 Konstanz 1, Germany
*Library of the Hungarian Academy of Sciences,
Information Science and Scientometrics Research Unit (ISSRU),
P.O. Box 7, H-1361 Budapest, Hungary

Abstract—A new measure of the citation speed of scientific publications based on a stopping time approach is proposed. A short theoretical introduction shows the mathematical exactness of the applied methods. The citation rate for papers published in 1980 has been recorded in the period 1980 through 1989 in three science fields. A sequence of the ith Harmonic Mean Response Times (i.e., the harmonic means of the time elapsed between the publication date and the date of the ith ($i = 1,2,\ldots$) citation of the papers) is analysed. The sequences can be approximated by linear functions. This empirical rule leads to the assumption that the complete citation succession and the rapidness of the ith response are already determined by the mean of the first response.

INTRODUCTION

The stochastic process approach is recently one of the most versatile tools in comprehending and modelling bibliometric phenomena. The growing portion in the bibliometric literature of the second half of the 1980s reflects its increasing importance, both for theoretical considerations and for practical use, such as constructing indicators or predicting frequencies. In particular, the changing impact of scientific information as time elapses since a scientific paper has been published and the change of circulations of monographs and scientific journals in libraries as a function of time form the major topics of the published analyses. The main results are above all connected with two names. Sichel (e.g., 1985) has applied his Generalized Inverse Gaussian-Poisson Distribution (GIGP) model to different bibliometric problems (author and journal productivity, citation rates, journal use, etc.). His distribution has three free parameters, one of which depends on time. The model is therefore appropriate to reflect phenomena subject to the lapse of time. Burrell (1988) has used mixtures of Poisson processes and a negative binomial process (1990) to analyse predictive aspects of some bibliometric processes. Recently, Glänzel and Schubert (1991) have used a nonhomogeneous birth process to show the predictability of mean citation rates and citation frequencies. A further interesting aspect that was not yet analysed is the succession (e.g., of citations to a given publication set) of papers published by a group of authors and of circulations in a library. Especially citation and circulation successions are of great practical importance. A first attempt to define an indicator connected with citation succession was the Mean Response Time (MRT) by Schubert and Glänzel (1986). The indicator, an exponential mean of the time of the first citation a set of papers receives, was originally formulated as an immediacy measure for journal citation speed. The indicator was applied to physics journals. In this paper we want to extend these previous results by defining an indicator measuring the succession speed as reflected by the ith response. The empirical properties of the indicator sequences will be demonstrated by several examples. Beforehand, a short theoretical introduction into special random variables in connection with stochastic processes is, however, indispensible.

*Permanent address.

1. THEORY

1.1 *A concise course on stopping times*

Consider a set of papers published in certain scientific journals during a given time period. We assume that this time period is short enough to be considered a single point s on the real-time scale. Without the loss of generality we put $s = 0$. Let $X(t)$ $(t \geq 0)$ denote the number of citations the papers in question receive during the time span t after their dates of publication. The (random) number of citations $\{X(t)\}_{t\geq 0}$ can then be considered a stochastic process (cf. Glänzel & Schubert, 1991). Under the assumption of particular underlying models this approach yields revealing results, concerning for instance certain predictive aspects (cf. Burrell, 1988, 1990; Glänzel & Schubert, 1991) or the estimation of annual citation rates (cf. Sichel, 1985). We now take up the random succession of citations the papers have received during a certain time span t. First of all we define some necessary tools. The process is assumed to be continuous in time. We further assume that the realizations of the process are continuous from the left (i.e., $X(t+) = X(t)$ almost everywhere). Let T_i denote the shortest time period t during which the papers have received exactly i citations (i.e., $T_i = \min\{t : X(t) \geq i\}F)$. Random variables of this type are called "stopping times." The measurability of the events $T_i < t$ or $T_i \leq t$, respectively, for each $i > 0$ and every $t \in [0,\infty]$ is obvious. In particular we have

$$P(T_i < t) = P(X(t-) \geq i), \tag{1}$$

provided $t > 0$. Here two important properties should be stressed.

1. Due to the right-continuity property we have

$$P(T_i \leq t) = P(X(t+) \geq i) \tag{2}$$

for every $t < \infty$ and $i > 0$.

Hence we obtain

$$P(T_i = t) = P(X(t) \geq i) - P(X(t-) \geq i), \tag{3}$$

which may differ from zero only at the discontinuity points of the process $X(t)$.

2. T_i can take the value $+\infty$ with positive probability. In particular we have

$$P(T_i = +\infty) = P(X(\infty) < i), \tag{4}$$

which may differ from zero according to whether the limiting distribution $P(X(\infty) = i)$ $(i = 0,1,2,\ldots)$ of the process $X(t)$ is degenerated or not.

The second property reflects the fact that a certain portion of papers may remain uncited; in other words, the probability that a paper will never be cited can be greater than zero. Instead of the continuous process as defined above, a discrete version of it is much more convenient in practice, namely,

$$Y(t) = X([t]) = X(t_n), \tag{5}$$

where $t_n = [t]$ denotes the integer part of the real argument t. Therefore $Y(t_n) = X(t_n)$ holds for all non-negative integers. In practice, $t_0 = s$ is usually the year of publication, t_1, $t_2, \ldots$ is the sequence of the years subsequent to the publication date. According to the above definitions we have

$$Y(t_n+) = Y(t_n)$$

and

$$Y(t_n-) = Y(t_{n-1})$$

and consequently, T_i here takes non-negative integer values. Hence

$$P(T_i = t_n) = P(Y(t_n) \geq i) - P(Y(t_{n-1} \geq i) \tag{6}$$

follows. For reasons of completeness we define $T_0 := t_0$. Finally we will make one single correction for practical use. Though in the continuous case $P(X(0) > 0) = 0$, we assume in the discrete case $P(Y(t_0) > 0) \geq 0$. This is a consequence of the former assumption that $t_0 = s$ is not a single point, but a shorter time period which may, as mentioned above, include a complete calendar year. Therefore the probability that a paper may be cited during this year can be assumed to be positive. Thus we have

$$P(T_i = t_0) = P(Y(t_0) \geq i),$$

which is exactly 1 if $i = 0$ (cf. the above definition of T_0). Herewith we have introduced all theoretical tools based on which the following analysis can be performed.

1.2 *"Stopping time" indicators*

If we attempt to derive particular distribution formulae for the random variables T_i ($i > 0$) from the original distributions of the stochastic process $Y(t_n)$, we are faced with several technical difficulties.

The first technical problem arises when we calculate the probabilities $P(T_i = t_n)$ for $i > 0$ (compare eqns (3), (4), and (6)). The resulting formulae in most cases take the form of finite sums, and are thus not very "handy." Secondly, the time behaviour of the discretized process often cannot be expressed by simple closed formulae. We therefore renounce the assumption of any particular underlying model. Equation (6), however, yields a useful tool for determining the time when a paper receives its ith citation. Thus the empirical estimates of the probabilities $P(T_i = t_n)$ can be derived from the ordinary citation rate distributions based on observations during t_{n-1} and t_n. The first simple indicator which can be defined with the help of the stopping time approach is the *Mean Response Time* (MRT). MRT was introduced by Schubert and Glänzel (1986) as an indicator of citation immediacy. Since the response time of all papers having received no citations in the observation period is to be taken as infinite, any arithmetic average is unsuitable in forming a response indicator. Therefore MRT is defined as the "exponential average" of the time of the first citation. The indicator was based on an observation period of five years, starting with the publication year.

$$\begin{aligned}\mathrm{MRT} &= -\ln(f_0 + f_1 e^{-1} + f_2 e^{-2} + f_3 e^{-3} + f_4 e^{-4}) \\ &= -\ln(G_1(0) + G_1(1)(e^{-1} - 1) + G_1(2)(e^{-2} - e^{-1}) + G_1(3)(e^{-3} - e^{-2}) \\ &\quad + G_1\,(4)e^{-4}),\end{aligned}$$

where f_t is the relative frequency of papers receiving their first citation in the tth year, and $G_1(t)$ is the portion of papers that have received at least one citation during the first t years from their dates of publication. This approximation does not cause considerable errors in any regard. Firstly, the probability of being cited for the first time later than five years after publication is very small; secondly, with regard to aspects of immediacy, it is quite indifferent to have received the first citation after 10, 25 years, or not at all. Thirdly, the exponential formula expresses these two arguments by decreasing the weight of latecomers:

$$e^{-t} \approx e^{-\infty} = 0 \qquad \text{for } t \geq 5.$$

It can be readily shown that MRT is an asymptotically unbiased estimator of the logarithm of the exponential expectation, that is:

$$-\ln E(\exp(-T_1)),$$

where T_1 is the stopping time as defined above. In order to uncover an interesting property of response and citation succession, we first of all define the ith Harmonic Mean Response Time (HMRT_i) as the harmonic mean of the ith response:

$$\mathrm{HMRT}_i = \left\{ \sum_{t=1}^{n} \frac{G_i(t) - G_i(t-1)}{t+1} + G_i(0) \right\}^{-1} - 1,$$

where $G_i(t)$ is the relative frequency of papers that have received at least i citations during time t, and n is a suitable maximum time period for observations. HMRT_i obviously estimates the expectation ratio

$$\frac{1}{E(1/(1+T_i))} - 1 = \frac{E(T_i/(1+T_i))}{E(1/(1+T_i))}.$$

As in the case of MRT, the transformation makes sure that HMRT_i takes values on the non-negative half of the real axis and that HMRT_i is an ascending function of the ith response (i.e., a greater HMRT_i value corresponds to a greater response time).

2. EXAMPLES

In order to illustrate the applicability of the above theory and to derive some empirical regularities, we use three samples we have already analyzed in connection with predictive aspects of citation processes (Glänzel & Schubert, 1991). All data used for the analysis were taken from the Corporate and Citation Index Files of the SCI® database of the Institute for Scientific Information (ISI, Philadelphia, PA, USA). All papers indicated as research articles, letters, notes, and reviews were taken into consideration. Papers published in 1980 were selected and all citations received by them in 1980 and the subsequent nine years have been counted. The following three sample source publications have been chosen:

1. all papers of the above type published in 1980 in the *Journal of the American Chemical Society* (*JACS*) representing the science field chemistry,
2. all papers published in 1980 dealing with problems of physics of condensed matter (CMPH) (i.e., in journals classified by Garfield into this Subject Category) (see, e.g., *SCI Annual Guide*, 1980), and
3. all papers published in 1980 in the journal *Lancet* representing the life sciences.

Tables 1–3 present the absolute frequencies of those papers that have received their ith citation in the tth ($t = 0, 1, \ldots, 9$) year after their publication. According to the above considerations, these frequencies coincide with the absolute frequencies of those papers that have received at least i citations during t years, but fewer than i citations still one year before.

Table 1. Absolute frequency distribution of the year of the ith response (T_i) for 1916 papers published in *JACS* in 1980

t	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
0	763	370	184	89	54	40	23	13	7	4
1	947	1061	938	791	641	473	389	306	230	189
2	157	350	518	628	650	681	642	597	557	501
3	22	66	127	159	221	269	287	319	329	322
4	5	18	45	76	102	135	162	176	193	216
5	4	10	27	44	69	69	88	98	114	118
6	3	2	10	18	25	41	54	58	76	97
7	1	8	11	10	18	34	36	47	51	55
8	3	5	5	9	13	13	30	39	44	51
9	0	1	2	6	11	11	23	28	24	21
>9	11	25	49	86	112	150	182	235	291	342

Table 2. Absolute frequency distribution of the year of the ith response (T_i) for 7414 papers concerned with physics of condensed matter and published in 1980

t	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
0	1345	416	146	61	23	12	4	3	1	0
1	2844	2102	1420	981	718	505	390	280	208	164
2	1215	1360	1240	1062	894	779	644	574	491	414
3	476	709	796	804	723	641	564	459	431	410
4	207	389	461	445	456	433	411	388	335	298
5	150	231	283	301	343	341	349	325	298	285
6	90	158	187	238	235	212	206	225	225	188
7	83	115	158	174	164	194	186	208	203	194
8	65	89	135	139	167	161	164	145	170	153
9	35	58	78	99	81	119	108	89	87	106
>9	904	1787	2510	3110	3610	4017	4388	4718	4965	5202

The relative frequency of $T_{10} = 9$ already changes between 0.5% (*JACS*) and 1.5% (CMPH). This may illustrate that statistical functions for T_i based on 10 years citation observation are not suitable for reliable studies if $i > 10$. In order to visualize this phenomenon we have contrasted the HMRT$_i$ values, which were calculated based on a 10-year citation period, with estimates determined based on a 5-year observation. Figures 1–3 show the results for the three samples. The deviation between each function pair grows with increasing i. Nevertheless, the *JACS* and *Lancet* data based on 10 years of observation seem to form an almost perfect straight line, starting from the origin of the coordinate system. The points (i, HMRT$_i$) of the CMPH data form a slightly convex curve which can, however, also be approximated by a straight line (see Fig. 4). Table 4 presents all HMRT$_i$ data ($i \le 10$) in particular.

Based on the data presented by Table 4 and Figs. 1–4, the formulation of the following empirical rule seems to be justified. HMRT$_i$ may be approximated by a linear function of the form $a \cdot i + b$ if i is small ($i \le 10$) and the observation period is great enough (10 years may be sufficient). Furthermore, b can be equated with 0, since $E(T_0/(1 + T_0)) = 0$ by definition. Thus we obtain

$$\mathrm{HMRT}_i \sim a \cdot i, \quad i = 0, 1, 2, \ldots,$$

where a is a positive real constant characterizing the response succession as reflected by citations. The reciprocal $1/a$ expresses the speed of the successive responses, and therefore will be called Average Rapidness of Citation Succession (ARCS). Table 5 presents the indicator values Mean Response Time, first Harmonic Mean Response Time, the average citation succession time (a), and ACRS (= $1/a$).

Table 3. Absolute frequency distribution of the year of the ith response (T_i) for 2286 papers published in *Lancet* in 1980

t	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
0	721	311	141	75	47	33	22	17	14	10
1	611	545	458	393	315	255	211	179	145	122
2	204	260	263	253	231	228	201	191	188	189
3	82	119	156	150	162	136	149	134	119	94
4	57	81	75	80	86	96	89	92	102	99
5	35	59	60	67	67	58	65	63	58	66
6	26	33	45	47	47	55	42	45	36	43
7	10	20	26	31	33	43	39	32	38	27
8	11	32	25	22	30	25	26	24	25	37
9	4	4	18	22	22	26	21	18	22	16
>9	525	822	1019	1146	1246	1331	1421	1491	1539	1583

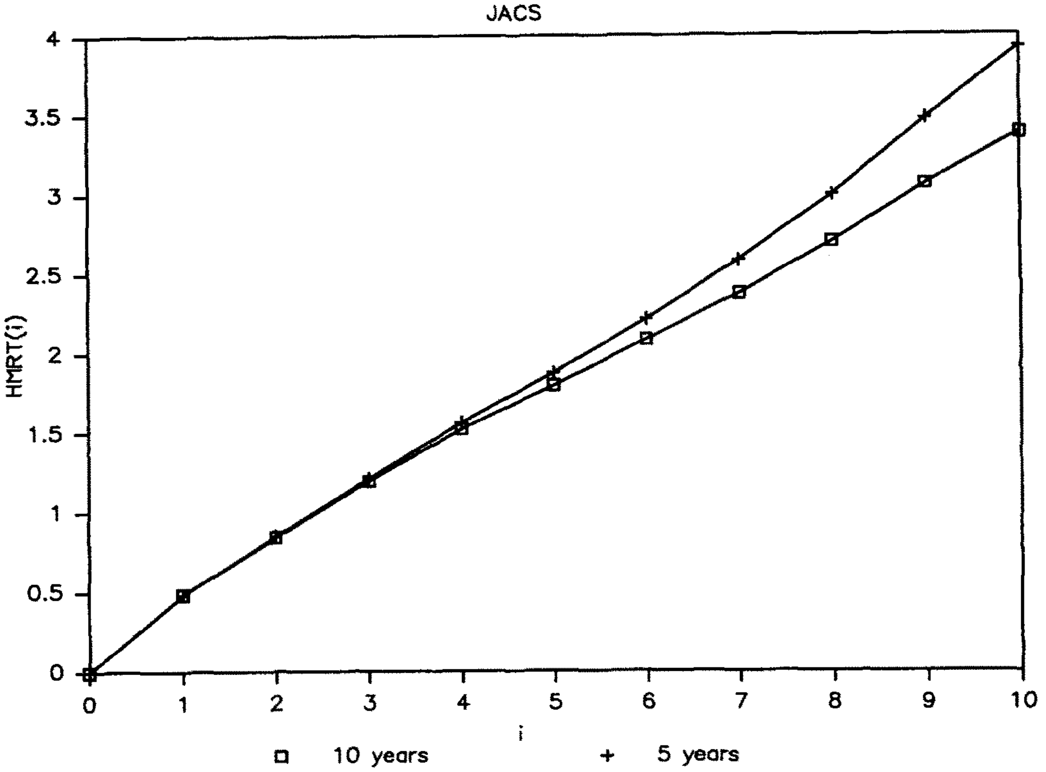

Fig. 1. Empirical $HMRT_i$ values based on five- and ten-year citation observations as a function of the ith response (*JACS* papers published in 1980).

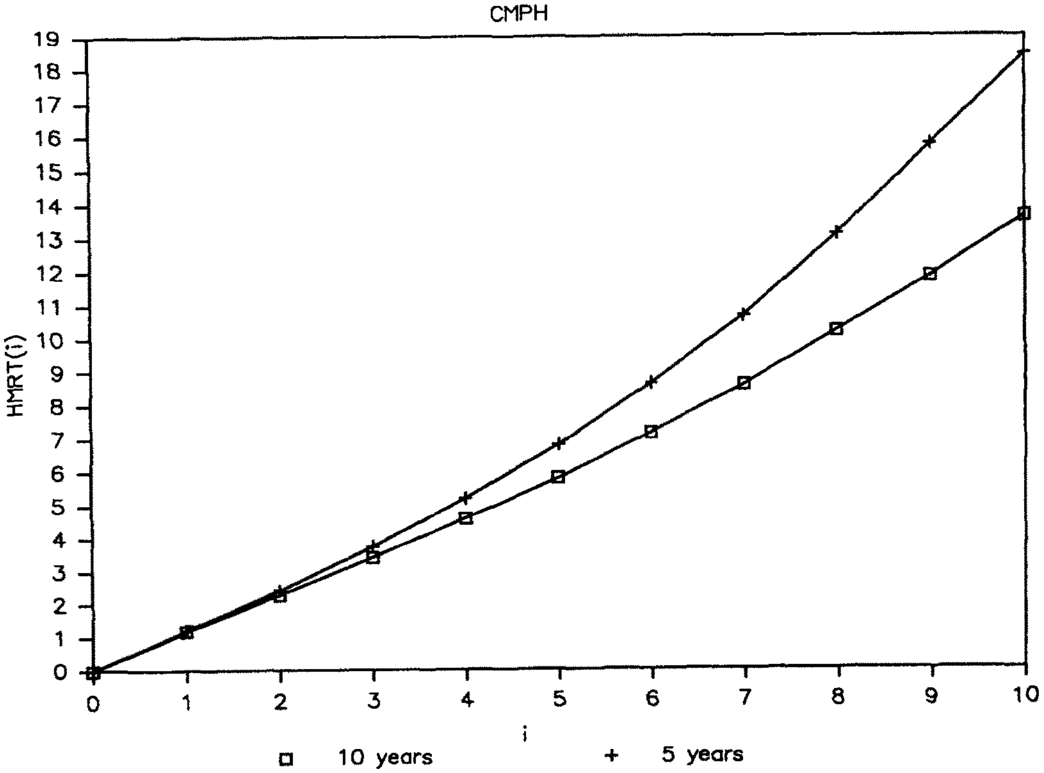

Fig. 2. Empirical $HMRT_i$ values based on five- and ten-year citation observations as a function of the ith response (CMPH papers published in 1980).

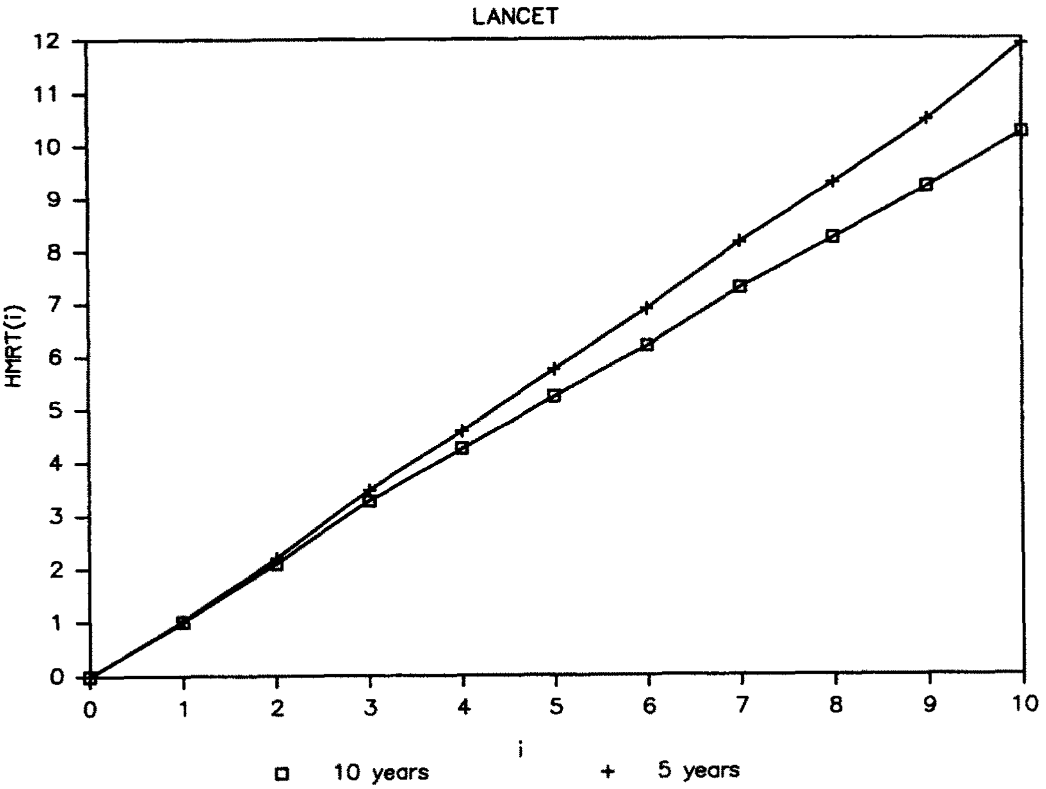

Fig. 3. Empirical HMRT$_i$ values based on five- and ten-year citation observations as a function of the ith response (*Lancet* papers published in 1980).

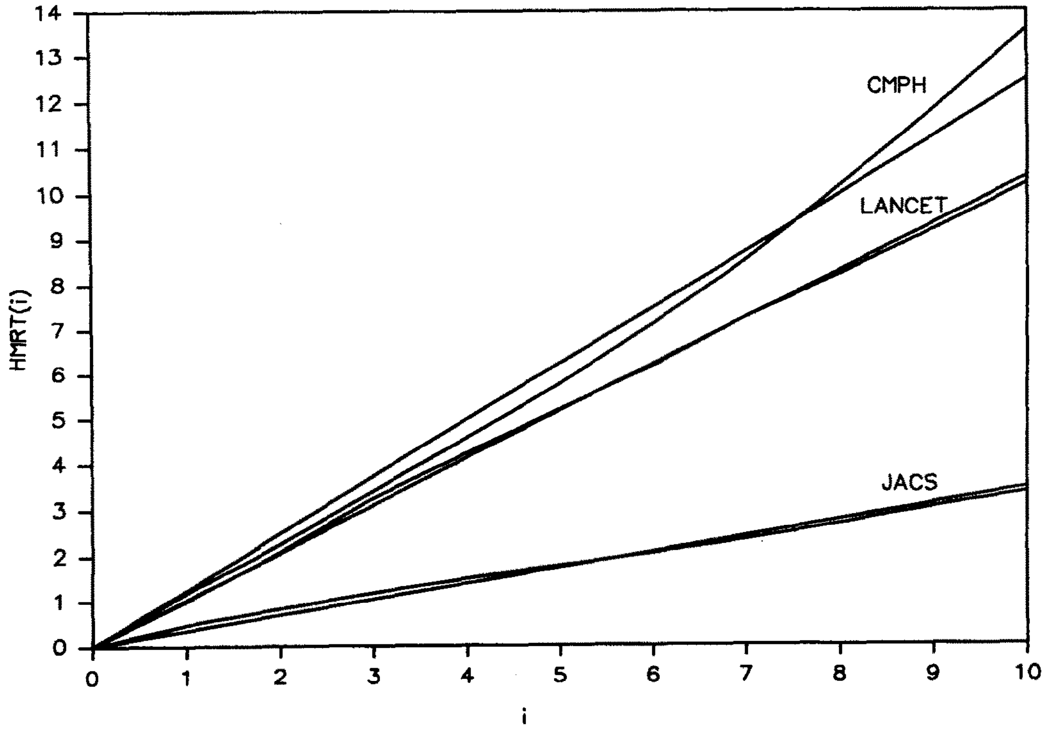

Fig. 4. Empirical and estimated HMRT$_i$ data based on ten-year observations for three samples.

Table 4. Empirical HMRT_i values based on five- and ten-year citation observations for all three samples

	JACS		*Lancet*		CMPH	
i	10a	5a	10a	5a	10a	5a
0	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1	0.48	0.48	1.01	1.03	1.19	1.22
2	0.84	0.85	2.10	2.19	2.27	2.41
3	1.19	1.21	3.27	3.47	3.42	3.75
4	1.52	1.56	4.25	4.58	4.59	5.20
5	1.79	1.87	5.23	5.74	5.80	6.79
6	2.08	2.21	6.19	6.89	7.13	8.62
7	2.37	2.58	7.28	8.16	8.55	10.64
8	2.70	2.99	8.24	9.29	10.17	13.08
9	3.06	3.47	9.21	10.47	11.82	15.72
10	3.38	3.92	10.23	11.89	13.59	18.40

Table 5. Indicators based on the stopping time approximation for three scientometric samples (based on 10 years citation observation)

	JACS	CMPH	*Lancet*
MRT	0.52	1.05	0.85
HMRT_i	0.48	1.19	1.01
a	0.35	1.25	1.04
ACSR	2.83	0.69	0.96

3. CONCLUSION

The somewhat surprising linearity behaviour of the harmonic mean response times permits the conclusion that the citation succession of scientific publications on different subjects is determined by the time of the first response. This contradicts any interpretation according to which the first citation cannot be considered to be significant, for example, because the first response is often a self-citation. The observed empirical rule concerning the HMRT_i sequences guarantees that the HMRT_1 (and as a consequence the MRT, too) is an "immediacy" indicator indeed. As already stressed by Schubert & Glänzel (1986), an (H)MRT value less than 1 reflects a rapid response, whereas a (H)MRT value greater than 1 corresponds to a slower one. The inverse relations hold for the ACRS. Thus the journal *Lancet* ranks with "fast" journals, and the journal *JACS* has an extremely high immediacy. The moderate immediacy of the papers published in physics of condensed matter is obviously due to the heterogenity of the subject field.

Acknowledgement—This research is in part supported by an Alexander von Humboldt Research Fellowship. The author wishes to acknowledge this support.

REFERENCES

Burrell, Q.L. (1988). Predictive aspects of some bibliometric processes. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 43–63). Elsevier Science Publishers B.V.

Burrell, Q.L. (1990). Empirical prediction of library circulations based on negative binomial processes. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90* (pp. 57–64). Elsevier Science Publishers B.V.

Glänzel, W., & Schubert, A. (1991). Predictive aspects of a stochastic model for citation processes. Unpublished manuscript.

Schubert, A., & Glänzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, *6*(3), 149–167.

Schubert, A., & Glänzel, W. (1986). Mean Response Time—A new indicator of journal citation speed with application to physics journals. *Czech. J. Phys.*, *B 36*, 121–125.

Sichel, H.S. (1985). A bibliometric distribution which really works. *JASIS*, *36*(5), 314–321.