Pergamon

Information Processing & Management, Vol. 31, No. 1, pp. 69-80, 1995

0306-4573/95 $9.50 + .00

0306-4573(94)E0012-Q

PREDICTIVE ASPECTS OF A STOCHASTIC MODEL FOR CITATION PROCESSES

W. GLÄNZEL[1,2] and A. SCHUBERT[2]
[1]Universität Konstanz, Sozialwissenschaftliche Fakultat, Fachgruppe Psychologie, Pf. 5560, D-7750 Konstanz 1, F. R. Germany
[2]Library of the Hungarian Academy of Sciences, Information Science and Scientometrics Research Unit (ISSRU), P.O. Box 7, H-1361 Budapest, Hungary

(*Received* 15 *August* 1993; *accepted in final form* 24 *March* 1994)

Abstract – A statistical model for citation processes is presented as a particular version of a nonhomogeneous birth process. The mean value function $E(X(t) - X(s)|X(s) = i)$ and special transition probabilities such as $P(X(t) - X(s) > 0|X(s) = 0)$ and $P(X(t) - X(s) = 0|X(s) > 0)$ give essential information on the change of citation impact in time. It is shown that the mean value functions and transition probabilities can readily be calculated on the basis of known and estimated parameters. The analysis is illustrated by five examples. The citation rate for papers published in 1980 has been recorded in the period 1980 through 1989 in five science fields. The model provides sufficiently good approximations for both the empirical mean value functions and the transition frequencies for the years 1985 and 1989 based on the number of citations the papers have received until 1982.

1. INTRODUCTION

The history of application of stochastic methods to bibliometrics and scientometrics is almost as long as the history of these specialties themselves. Although at the very beginning the detection and formulation of fundamental regularities regarding bibliometric phenomena such as the publication activity of scientists, the frequency of circulation in libraries, and citation frequencies of scientific publications were in the limelight, growing interest has recently been focussed on comprehending the mechanism of these phenomena and on modeling their change in time. In particular, bibliometricians are nowadays not only interested in knowing the distribution of papers researchers have published, circulations that monographs and scientific journals have, and citations that different papers receive, but also in understanding and predicting, for example, the consequences on the productivity of a scientist if he or she becomes more experienced but less ambitious, or how the number of citations received reflects the changing impact of scientific information as time elapses since the paper was published. Thus, the concept of the stochastic process has been adopted by the bibliometric literature, too. Sichel (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, particularly appropriate to reflect explicitly time-dependent phenomena. Burrell (1990) has used mixtures of Poisson processes to analyse predictive aspects of some bibliometric processes. In particular, beside the application of the Gamma-Poisson and the GIGP process, he has constructed the Generalized Waring process in his paper. Finally, we would like to mention the dynamic Waring model by Schubert and Glänzel (1984), the limiting distribution of which has successfully been used to characterize publication activity of authors (Schubert & Glänzel, 1984b; Boxenbaum *et al.*, 1987). In what follows, a simple model closely related

This is a new version of a paper presented at the Third International Conference on Informetrics in Bangalore (India), August 9-12, 1991.

to this latter one will be built for the citation process. The predictive power of the model will be demonstrated by several examples.

2. THEORY

2.1 *The stochastic model*

The dynamic Waring model introduced by Schubert and Glänzel (1984a) will serve as a general basis of the stochastic model of citation processes, too. We shall, therefore, first recall this model in brief. Thereafter, we shall use a modified version of this model to reflect the specific mechanism of citation processes.

2.1.1 *The dynamic Waring model.* Consider an infinite array of units indexed in succession by the non-negative integers. The content of the ith unit is denoted by x_i, the (finite) content of all units by x. Then the fraction $y_i = x_i/x$ $(i \geq 0)$ expresses the share of elements contained by the ith cell. The change of content is postulated to obey the following rules.

1. Substance may enter the system from the external environment through 0th unit at a rate s.
2. Substance may be transferred unidirectionally from the ith unit to the $(i + 1)$th one at a rate $f_i (i \in N_0)$.
3. Substance may leak out from the ith unit into the external environment at a rate g_i $(i \in N_0)$.

The next step towards a stochastic model is to interpret the above ratios y_i as the (classical) probability with which an element is contained by the ith unit. The stochastic process is then formed by the change of the content of the units (i.e., by the change of papers published by the authors who have entered the system). $X(t)$ denotes the (random) number of published papers, $P(X(t) = i) = y_i$ the probability that an author in the system has published exactly i papers in the period t. Finally, the stochastic model itself is obtained if $X(t)$ is considered the *publication activity process* of an arbitrary author, and $P(X(t) = i) = y_i$ is the probability that he or she has published i papers in the interval $(0, t)$. Figure 1 visualizes the scheme of substance flow of this process.

Now, using the above notations, we can give a mathematical formulation for the equations of change in the system.

$$\begin{aligned} x_0'(t) &= s(t) - f_0(t) - g_0(t), \\ &\vdots \\ x_i'(t) &= f_{i-1}(t) - f_i(t) - g_i(t), \quad (i > 0). \end{aligned} \tag{1}$$

Here and in the following, the prime (′) denotes time derivatives.

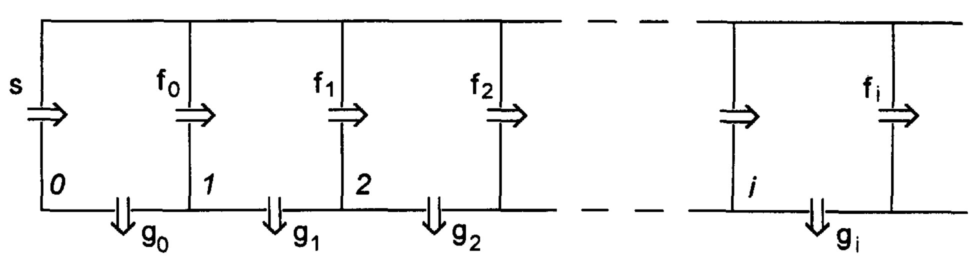

Fig. 1. Scheme of substance flow in the Waring process.

According to Schubert and Glänzel, the following particular forms of the above rate terms are used:

$$s = \sigma \cdot x, \tag{2}$$

$$f_i = (a + b \cdot i) \cdot x_i; \quad (i \geq 0), \tag{3}$$

$$g_i = \gamma \cdot x_i; \quad (i \geq 0), \tag{4}$$

where σ, a, b, and g are non-negative real values. Since $x'(t) = \sum x_i' = (s - \sum g_i) = (s - \gamma)x$ (cf. eqn (1)), the distribution of the substance over the units during time t can be obtained as a solution of the following system of first-order linear differential equations:

$$\begin{aligned} y_0'(t) &= \sigma - (a + \sigma) \cdot y_0 \\ &\vdots \\ y_i(t) &= (a + b \cdot (i - 1) \cdot y_{i-1} - (a + b \cdot i + \sigma) \cdot y_i; \quad i > 0, \end{aligned} \tag{5}$$

with the initial conditions

$$y_i(0) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For the entire population we can derive $x(t) = x(0) \cdot \exp((\sigma - \gamma) \cdot t)$; that is, the system is asymptotically time-invariant (stationary) if $\sigma = \gamma$, otherwise, if $\sigma > \gamma$ or $\sigma < \gamma$, it exponentially grows or decays, respectively. The general solution of the above system of differential equations is

$$y_i = \sum_{j=0}^{i} b_{ij} e^{-(a+b\cdot i+\sigma)t} + \frac{\sigma(a + b) \ldots (a + b(i - 1))}{(a + \sigma)(a + b + \sigma) \ldots (a + bi + \sigma)}, \tag{6}$$

where the coefficients b_{ij} are determined by the initial conditions. Finally, if $\sigma = \gamma$ is assumed, the above process has a nondegenerated limit distribution, which is a Waring distribution with parameters $N = a/b$ and $\alpha = \sigma/b$.

As the model suggests, the Waring distribution can be considered a reasonably good approximation for finite times large enough, when the substance can be supposed to reach a distribution sufficiently close to the limiting case. Otherwise, if t is small, the distribution is less "handy" (see eqn (6)).

Next we shall consider a model that does not assume any interaction with the external environment.

2.1.2 *A nonhomogeneous birth process.* Consider now the same array of units as above. In contrast to the above model, we now assume that the system is completely isolated from external influences (i.e., no substance enters or leaves the system). Therefore, only rule (2) of the preceding section remains valid. Hence, $x(t) = x(0)$ follows immediately, where, of course, $x(0) > 0$ is assumed. The special assumption $x(0) = 1$ does not mean any restriction of generality. Now we reformulate eqns (2)–(4):

$$\sigma = 0, \tag{2'}$$

$$f_i = (a + bi) \cdot x_i; \quad a(t)/b(t) = \text{const } (>0), \tag{3'}$$

$$g_i = 0; \quad (i \geq 0). \tag{4'}$$

The subsidiary condition to eqn (3′) says that the process is nonhomogeneous (i.e., the substance flow may depend on the time elapsed). The proportionality coefficient in the subsidiary condition is denoted by N (i.e., $a(t) = N \cdot b(t)$). Figure 2 shows the scheme of substance flow of this process.

By analogy to the Waring model, we have

$$\begin{aligned} y_0'(t) &= -N \cdot y_0 \cdot b(t) \\ y_1'(t) &= (N \cdot y_0(t) - (N+1)y_1(t)) \cdot b(t) \\ &\vdots \\ y_i'(t) &= ((N+i-1) \cdot y_{i-1}(t) - (N+i)y_i(t)) \cdot b(t), \end{aligned} \qquad (5')$$

with the same initial conditions as above.

Publication-activity dynamics is basically reflected by the distribution $P(X(t) = i)$ of the process. Special conditional probabilities, the so-called transition probabilities, however, permit a much deeper insight into scientific productivity processes. In our case, for example, the probability that an author publishes j papers during t years, provided he or she has already published i papers during s years, is called transition probability ($i \le j$, $s < t$). These probabilities reflect the influence of an initial period on the further publication activity. In particular, the probability that at time t the substance is in the kth unit, provided it was in the ith one at time s, is denoted by $p_{ik}(s,t)$ ($k \ge i$). With the transition rules as above, we can write (see, e.g., Karlin & Taylor, 1975)

$$\partial p_{ik}(s,t)/\partial t = \{(N+k-1)p_{ik-1}(s,t) - (N+k)p_{ik}(s,t)\} \cdot b(t); \quad k > i. \qquad (7)$$

The initial conditions are

$$p_{ik}(s,s) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, let $X(t)$ denote the (random) index of that unit in which the substance is at time t. Then the following distribution can be obtained from the first system of differential equations by successive integration

$$P(X(t) = k) = \binom{N+k-1}{k} e^{-r(t)N}(1 - e^{-r(t)})^k, \qquad (8)$$

where $r(t) = \int_0^t b(u)\,du$ (i.e., the distribution of substance over the units is negative binomial at any time). Analogously, the second system of differential equations results in

$$\begin{aligned} p_{ij}^*(s,t) &= P(X(t) - X(s) = j \mid X(s) = i) \\ &= \binom{N+i+j-1}{j} \cdot e^{-(r(t)-r(s))(N+i)}(1 - e^{-(r(t)-r(s))})^j, \end{aligned} \qquad (9)$$

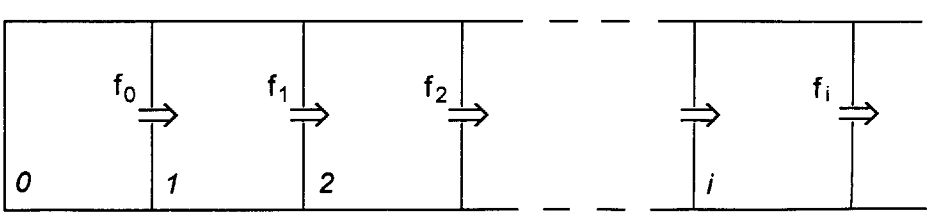

Fig. 2. Scheme of substance flow in the nonhomogeneous birth process.

where $p^*_{ij}(s,t) = p_{ik}(s,t)$ with $i = k - j \geq 0$. Equation (9) can be reformulated verbally as follows. The substance flow during the time period t–s has a negative binomial distribution with parameters $\exp(-r(t) + r(s))$ and $N + j$, where j is the index of the unit reached by the substance at time s.

The mean value functions of the process are defined as

$$M_i(s,t) = E(X(t) - X(s)|X(s) = i). \tag{10}$$

This function will play an important role in the applications. Under the above conditions we have

$$M_i(s,t) = (N + i)(\exp(r(t) - r(s)) - 1); \quad i \geq 0, t \geq s \tag{11}$$

and

$$M(s,t) = E(X(t) - X(s)) = N \cdot (\exp(r(t)) - \exp(r(s))). \tag{12}$$

Equation (12) reflects the nonhomogeneity of the process; for example, $M(s,s+h) \neq M(t,t+h)$ if $s \neq t (h > 0)$. The process has a nondegenerated or degenerated limit distribution as $\lim r(t) < +\infty$ or $\lim r(t) = +\infty$, respectively.

3. APPLICATIONS

3.1 *General remarks*

Both models above can usefully be applied in bibliometrics, each describing different phenomena. Let the main differences in their applications be outlined first. Consider an open population of scientists publishing papers regularly, say, in a given field of science. Newcomer authors enter the system through the 0th unit, and terminators are leaving it through the unit representing their highest productivity. The transition intensity follows the cumulative advantage principle. The dynamic Waring model seems to be an appropriate model to describe such a scientific community. The interaction with the external environment is obvious. If, on the other hand, a set of papers published within a certain time period, say, in a given year, is considered and the number of citations received by them in the subsequent years are counted, then the second model meets all formal requirements. The size of the set is *a priori* fixed; new publications do not enter the system, and papers do not leave it, either. The citation process requires a model that supplies a nondegenerate limit distribution. Citation processes are obviously nonhomogeneous.

One of the most interesting fields of application is the prediction of events, frequencies, and expectations. Thus, eqn (12) states that the expected number of citations received within a time span (t–s) is a linear function of the number of all citations that the paper(s) in question had received previously. The following results could also help to extrapolate citation-based indicators to longer time periods.

3.2 *Five examples*

In order to illustrate the applicability of the theoretical considerations above to empirical citation data, we have chosen five particular samples representing five different areas of science. Since it is practically impossible to compile lifetime citation data, only citations received during the first 10 years after the date of publication were counted. Within the limits of this restriction we will show that our model allows valid extrapolations and predictions. All data used for the analysis were taken from the Corporate and Citation Index Files of the *Science Citation Index*® (SCI) and *Social Science Citation Index*® (SSCI) databases of the Institute for Scientific Information (ISI, Philadelphia, PA). 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. If the sample size has been too small for reliable analysis data for the time period, 1981–1990 have been added to the set in analogous matter. The following five subject fields have been chosen:

1. chemistry,
2. condensed matter physics,
3. general and internal medicine,
4. psychology, and
5. probability theory and mathematical statistics.

The above science areas are represented by the following samples of source publications:

1. all papers published in 1980 in the *Journal of the American Chemical Society* (JACS, 1916 papers),
2. all papers published in 1980 in the subfield of condensed matter physics, that is, in journals classified in the SCI into this subject category (CMPH, 7414 papers),
3. all papers published in 1980 in the journal *Lancet* (LANCET, 2286 papers),
4. all papers published in 1980 and 1981 in the journal *Developmental Psychology* (DP, 203 papers), and
5. all papers published in 1980 and 1981 in the journal *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* (since 1985, *Probability Theory and Related Fields*) (ZWVG, 223 papers).

According to our model, the annual citation rates of each process have to obey a negative binomial distribution with one constant parameter (N) and one parameter $q(t)$, which may depend on time. This property of $q(t)$ is reflected by the increments in Figs. 3 to 7, which illustrate that the citation processes are nonhomogeneous, indeed. The parameters of the distribution can be estimated by the help of the maximum likelihood estimation or the method of moments. Unfortunately, both methods tend to fail in case of social processes such as bibliometric distributions, since the outstanding performance of a very small "elite" used to differ significantly from those that were expected on the basis of the performance of all other members of the population (see Glänzel & Schubert, 1988). This

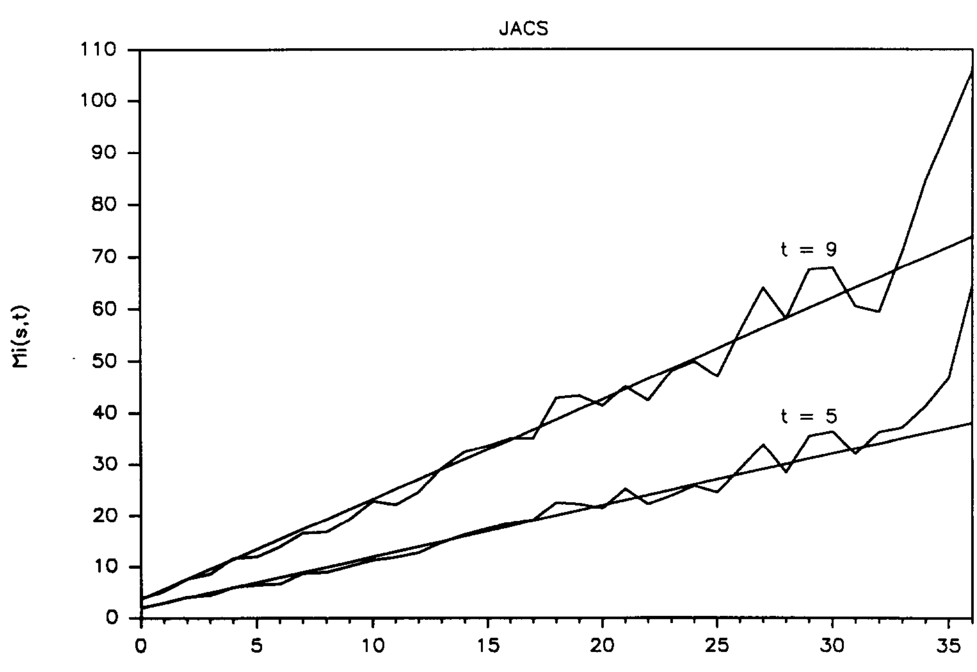

Fig. 3. Empirical and estimated values of the mean value functions $M_i(s,t)$ with $s = 1982$, $t = 1985$ and 1989 for papers published in *JACS* (1980).

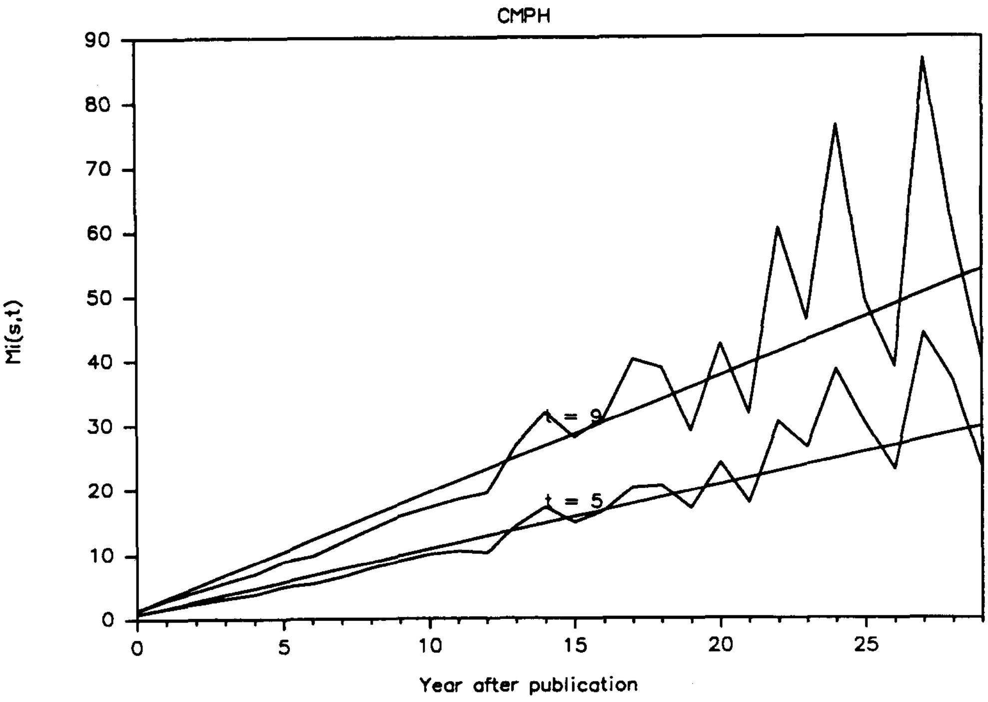

Fig. 4. Empirical and estimated values of the mean value functions $M_i(s,t)$ with $s = 1982$, $t = 1985$ and 1989 for papers published in 1980 and concerned with physics of condensed matter.

effect may cause heavy distortions for the variance estimate, as well as in case of the ML estimates. Therefore, we have used an estimation method using the sample mean and the fraction of uncited papers. This method was already used by Schubert and Glänzel (1983) in order to obtain less distorted standard deviation estimates for negative binomial distri-

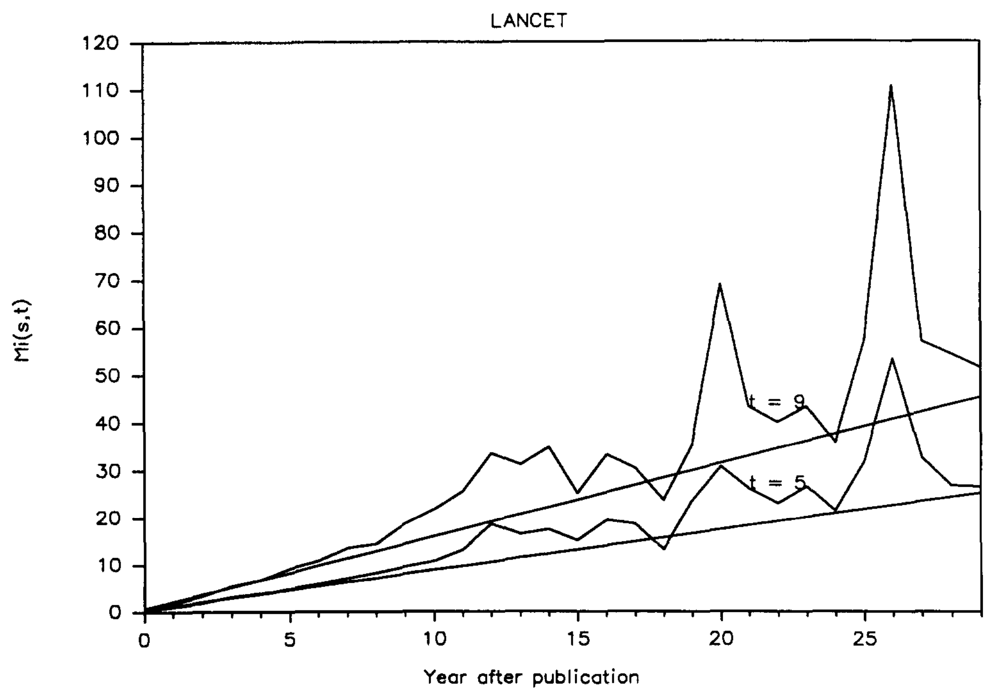

Fig. 5. Empirical and estimated values of the mean value functions $M_i(s,t)$ with $s = 1982$, $t = 1985$ and 1989 for papers published in *Lancet* (1980).

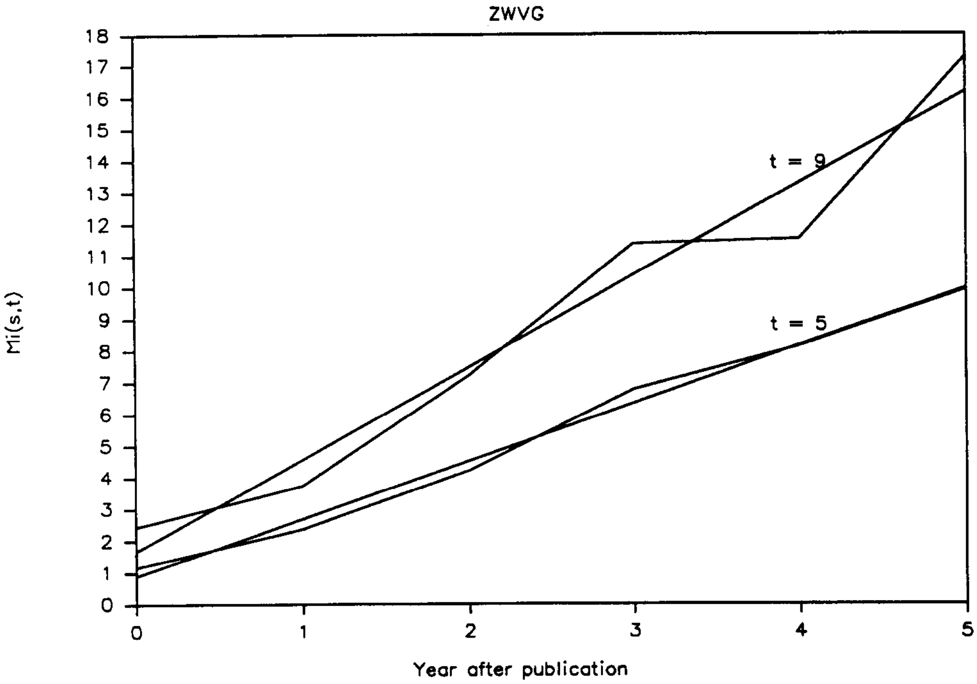

Fig. 6. Empirical and estimated values of the mean value functions $M_i(s,t)$ with s = 1982 (1983), t = 1985 (1986) and 1989 (1990) for papers published in *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* (1980–1981).

butions, and proved to be extremely stable, even if the fraction of uncited papers is small. We have

$$Q^*(t) = 1 - x(t)(\log Q^*(t)/(\log f_0(t))$$

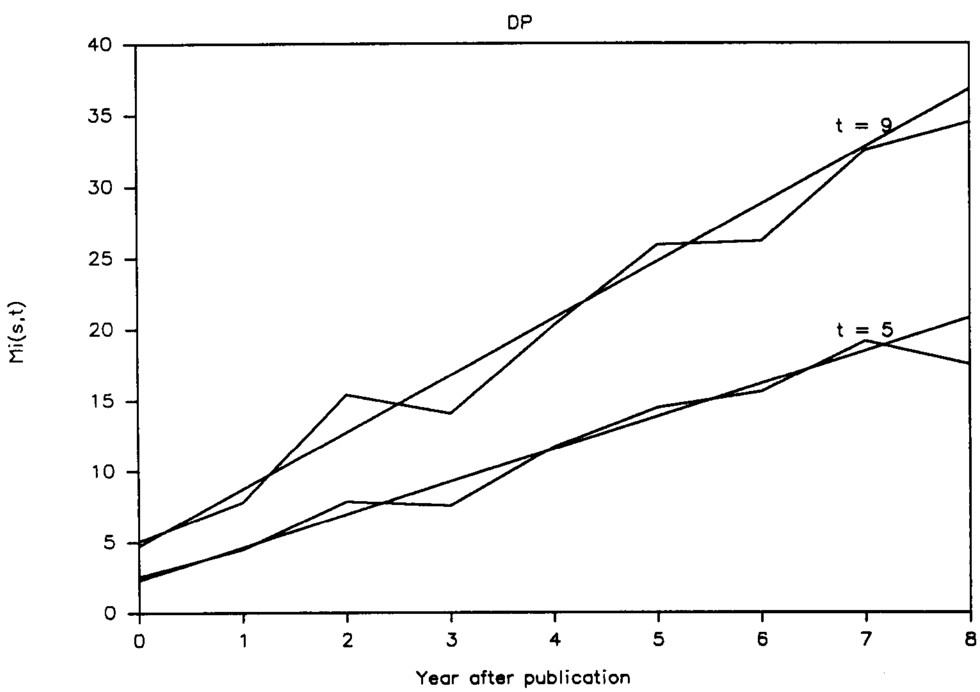

Fig. 7. Empirical and estimated values of the mean value functions $M_i(s,t)$ with s = 1982 (1983), t = 1985 (1986) and 1989 (1990) for papers published in *Developmental Psychology* (1980–1981).

and

$$N^* = x(t)/(1/q^*(t) - 1),$$

where N^* and $q^*(t) = 1/Q^*(t)$ are estimates of the parameters of the distribution, and $x(t)$ and $f_0(t)$ are the sample mean and the relative frequency of uncited papers at time t, respectively. The estimated parameter values for all five samples can be found in Table 1.

As the predictive aspects are regarded, the mean value function and certain transition probabilities play a particularly important part. We will estimate the probability and the expected values of future citation rates based on available data for an initial time period if the constant parameter N is known and if we have enough information about the change of substance flow in time (see above model). The function $r(t)$ in the exponent of the parameter $q(t) = \exp(-r(t))$ (cf. eqn (8)) is supposed to have one of two classical forms, namely an exponential or a hyperbolic one. Since we do not have any information about the limiting values $r(\infty)$, it is, however, quite impracticable to obtain reliable estimates of all parameters of the function $r(t)$ from nine empirical data each. Therefore, we use the estimates of $q(t)$ directly without caring about its particular mathematical form. From eqns (8), (9), and (12), we thus obtain the following two important equations:

$$P(X(t) - X(s) = k|X(s) = i) = \binom{N+i+k-1}{k} \cdot q(s,t)^{N+1}(1 - q(s,t))^k, \quad (13)$$

where

$$q(s,t) = (E(X(s)) + N)/(E(X(t)) + N)$$

and

$$M_i(s,t) = (N+i)(E(X(t) - X(s)))/(E(X(s) + N). \quad (14)$$

Equation (14) expresses the expected citation rate during the time period $(t - s)$ under the condition that the paper in question has received i citations during the time span s. With regard to the applications, this function may be the most important one.

Concerning eqn (13), two special probabilities are of particular interest:

$$P(X(t) - X(s) > 0|X(s) = 0) = 1 - q(s,t)^N; \quad (15)$$

that is, the probability that a hitherto uncited paper gets cited, and

$$P(X(t) - X(s) = 0|X(s) = i) = q(s,t)^{N+i}; \quad (16)$$

that is, the probability that a paper that has received $i (i \geq 0)$ citations will not be cited anymore (if $t \gg s$). Note that the latter probabilities form a strict geometric sequence.

3.2.1 *Choice of time parameters for observation and prediction.* When the time parameters for the observation and the prediction period are to be chosen, first of all the

Table 1. Estimated parameters of the mean value functions $M_i(s,t) = u \cdot i + v$ for $s = 1982$, $t = 1985$ and $t = 1989$, respectively $(v = u \cdot N)$

	JACS		CMPH		LANCET		DP		ZWVG	
	t_5	t_9	t_5	t_9	t_5	t_9	t_5	t_9	t_5	t_9
N	1.91		0.79		0.44		1.18		0.58	
u	1.00	1.95	1.00	1.81	0.86	1.54	2.30	4.01	1.81	2.90
v	1.92	3.71	0.79	1.43	0.38	0.68	2.34	4.73	0.88	1.67

fact should be taken into consideration that neither eqns (15) nor (16) would yield relevant information for arbitrary time parameters s and t. Whereas eqn (15) gives interesting information if s is rather small, eqn (16) should be applied if the observation period s already covers a greater time span and t tends to infinity. Equations (15) and (16) are therefore rather designed for alternative use. Equation (14), however, gives relevant information for any pair of time parameters (s,t). In case of citation processes, choosing a rather small observation period s and an arbitrary prediction time t may obviously be reasonable. On the other hand, one should resist any temptation to extrapolate data based on a too small observation period, since the standard deviation grows rapidly as s tends to zero. Thus, the standard deviation of the estimated values of the mean value function $M_i^*(s,t)$ can be approximated by

$$D(M_i^*(s,t)) \approx [(1 + i/N)D(X(t) - X(s))]^{1/2}/[nP(X(s) = i)]^{1/2}, \tag{17}$$

where the denominator almost vanishes if s is close to zero and i is positive. By theoretical and practical reasons, an observation period of two years seems reasonable. For our analysis we have therefore chosen the time period 1980–1982 ($s = t_2$) for observations and the years $t_5 = 1985$ and $t_9 = 1989$ for predictions. The mean value functions have the forms $M_i(s,t_5) = u_5 \cdot i + v_5$ and $M_i(s,t_9) = u_9 \cdot i + v_9$, respectively. The parameters u and v were calculated based on eqn (14), and can be found in Table 1. The empirical means were calculated as the average citation rates during 1983–1985 and 1983–1989, provided the papers had received up to 1982 $i = 0,1,2,\ldots$ citations. Empirical means and estimated mean function values for JACS, CMPH, LANCET, DP, and ZWVG papers are presented in Figs. 3 to 7. Data for $i \geq 30$ are not shown, since absolute frequency values start to be zero at $i = 30$, and therefore the empirical conditional means are not defined at this point. According to eqn (17), the standard deviation of the estimated conditional means grows rapidly with increasing argument i, since the numerator is a strictly ascending function and the denominator a descending function of i, if i is great enough. This effect can be observed in all five figures.

Finally, we would like to have a look at the transition probabilities $P(X(t) - X(s) = k \mid X(s) = 0)$. For the analysis, all papers of the above five samples were taken into consideration; they remained uncited during the first three years. We have arranged the observed citation rates into frequency distributions according to how many citations these papers received during 1983 and 1989. The corresponding theoretical values were estimated based on the parameter values presented in Table 1. Table 2 shows the results. The great chi-squared figures in case of the CMPH and ZWVG are obviously caused by the unexpected long tail of the distribution, since the fit is otherwise absolutely acceptable. According to eqn (15) we have in particular: 1106 out of 2010 uncited CMPH papers were cited during 1983 and 1989 (estimated value 1093), 38 out of 49 uncited JACS papers were cited during the same time period (estimated value 43), 225 out of the 750 uncited LANCET papers received citation after 1982 (estimated value 283), 32 out of the 35 uncited DP papers received citations after 1982 (estimated value 30), and 77 out of 109 uncited ZWVG papers were cited during 1983 and 1989 (estimated valued 59).

4. DISCUSSION

The above calculations are not real predictions in the sense of forecasting probabilities and expectations of future events, since estimates of the parameters $q(s,t)$ based on five and ten years' observations were used. Nevertheless, some important conclusions can be derived for possible medium-term forecasts. The JACS sample shows the "most correct" behaviour. LANCET allows the least reliable predictions. The two samples from hard science fields seem to represent similar regularities regarding the time parameter $u(t)$ (cf. eqn (5′) and Table 1). u_{1985} is very close to the value 1, and u_{1989} is roughly equal to 1.9 in either case. The different constant parameter N, however, causes completely different distributions. While the annual distributions of the citation rate of CMPH papers remain extremely skewed for all $t = 0,\ldots,9$, the peak of the annual citation distributions

Table 2. Absolute frequencies (*obs.*), estimated transition probabilities (*est.*) of the citation rate distribution of previously uncited papers with a χ^2 test and critical values at a confidence level at 95% and 99%

	JACS		CMPH		LANCET		DP		ZWVG	
k	obs.	est.	obs.	est.	obs.	est.	obs.	est.	obs.	est.
0	11	6.2	904	889.9	525	496.0	3	5.2	32	49.7
1	9	7.9	487	451.9	110	133.4	7	4.9	35	21.3
2	4	7.5	245	260.3	54	58.4	4	4.3	11	12.5
3	7	6.5	137	155.3	26	28.8	4	3.7	7	8.0
4	4	5.3	90	95.1	12	15.0	4	3.1	7	5.3
5	3	4.1	46	58.7	6	8.1	0	2.5	4	3.6
6	2	3.1	32	36.5	4	4.5	2	2.1	3	2.5
7	4	2.3	24	22.8	5	2.5	1	1.7	2	1.7
8	1	1.7	12	14.3	1	1.4	2	1.4	0	1.2
9	0	1.2	13	9.0	3	0.8	2	1.1	1	0.8
10	1	0.9	2	5.7	2	0.5	1	0.9	1	0.6
>10	3	2.2	18	9.8	2	0.5	5	4.0	6	1.6
Uncited 80-82	49		2010		750		35		109	
Sample mean	3.71		1.43		0.68		4.74		2.45	
Total size	1916		7414		2286		203		223	
Uncited 80-89	0.6%		12.2%		23.0%		1.5%		13.3%	
Total mean	31.5		10.28		14.0		15.0		5.9	
χ^2	6.58		42.82		17.61		6.14		19.26	
Dg.f.	6		14		7		9		6	
95%	12.6		23.7		14.1		16.9		12.6	
99%	16.8		29.1		18.5		21.7		16.8	

of the JACS papers differs from 0 (cf. total mean and portion of total uncited papers in Table 2). A further consequence is the difference in the transition probabilities. Anyway, it is remarkable that more than three quarters (78%) of all JACS papers uncited up to 1982 received citations during 1983 and 1989. In the subfield physics of condensed matter, this portion was somewhat smaller, but still greater than one half (55%). The citation rates of papers published in LANCET seems to form a completely different citation process. On one hand, the constant parameter N is strikingly small, expressed by the extremely skewed annual distributions (cf. the portion of uncited papers in Table 2). On the other hand the increments as well as the parameter $u(t)$ reflect the fact that processes here proceed much faster than in the case of the other examples. Further evidence is that only 30% of uncited papers have received citations during 1983 and 1989. This phenomenon may be caused partially by subject characteristics, but also by the fact that only a relatively small part of papers published in LANCET are research articles; a major part are letters. The regularities of citation processes are, however, well reflected by the chosen model in this case, too. The two "small" samples representing mathematical statistics and psychology show surprisingly correct behaviour. The mathematical journal behaves more like a social science journal than a natural science one. The u parameters are in both cases much greater than in the case of the other three journals. The parameters of the DP sample (cf. Table 1) show that the shape of the annual distribution resembles that of JACS. As in the case of JACS, the mean citation rate of those papers uncited during the first three years is remarkably great (4.74). On the other hand, the small parameter N of the ZWVG sample reflects much more skewed annual distributions. Finally, we wish to note that an interesting application field of the above mentioned model may be the extrapolation of citation-based indicators, which are often calculated for a time period of one or two years following the publication date. The mean value function may then help to estimate mean citation rates for longer or overlapping time periods.

Acknowledgement—This research is in part supported by an Alexander von Humboldt Research Fellowship. The first author wishes to acknowledge the AvH support. The authors wish to thank Urs Schoepflin (Max Planck Institute for Human Development and Education, Berlin) for providing the data concerning the journals *Developmental Psychology* and *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*.

REFERENCES

Boxenbaum, H., Pivinski, F., & Ruberg, S.J. (1987). Publication rates of pharmaceutical scientists: Application of the Waring distribution. *Drug Metabolism Reviews*, *18* (4), 553–571.

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. (1988). Theoretical and empirical studies on the tails of scientometric distributions. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 75–83). Elsevier Science Publishers B.V.

Glänzel, W., & Schubert, A. (1990). The cumulative advantage function. A mathematical formulation based on conditional expectations and its application to scientometric distributions. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90* (pp. 139–147). Elsevier Science Publishers B.V.

Karlin, S., & Taylor, H.M. (1975). *A first course in stochastic processes.* Academic Press, Inc.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, *5* (1), 59–74.

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

Schubert, A., & Glänzel, W. (1984b). Prompt nuclear analysis literature: A cumulative advantage approach. *Journal of Radioanalytical and Nuclear Chemistry, Articles*, *82* (1), 215–218.

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