ON THE POSSIBILITY AND RELIABILITY OF PREDICTIONS
BASED ON STOCHASTIC CITATION PROCESSES

W. GLÄNZEL*

Research Association for Science Communication and Information e.V. (RASCi),
Johannes-Kepler-Weg 5, D-15236 Frankfurt (Oder) (Germany)

A statistical model for citation processes, a particular version of a non-homogeneous birth
process, is analysed in the context of predictions of future citation rates. Important properties of
the process were already studied by the author in earlier papers. Although the applicability of the
model was demonstrated by several examples, practical aspects of predictions and questions of
statistical reliability were not tackled so far. The present study is focused on the demonstration of
the possibility of true predictions and on the analysis of the statistical reliability of predictions
based on the mean value function $E(X(t) - X(s) | X(s) = i)$ of citation processes. The citation rates for
papers published in 1980 and 1991 were recorded in the period 1980 through 1995, and 1991
through 1995, respectively, in all science areas. It is shown that parameters of mean value
functions estimated for earlier time periods can be applied to more recent years, too. As a by-
product, the model may serve as a validation tool for the particular choice of citation windows in
evaluation studies.

1. Introduction

In some recent papers, predictive aspects of bibliometric processes were analysed in
the context of information use, that is, in the traditional approach models for library
circulations were introduced, studied and discussed. The first complete model was
developed by Burrell (1988, 1990) and Burrell and Fenton (1992, 1994). Burrell and
Fenton used mixtures of Poisson processes to describe frequency-of-circulation
distributions as a function of time and to predict future use of documents. The later
versions of the model Burrell and Fenton (1992, 1994) also incorporated loan periods.
Although the model was designed for application to information use in libraries, Burrell
pointed out that his model may be applied to other bibliometric processes such as
citation processes, too. The second complete model for the change of information use in
time in the context of possible predictions was introduced by Glänzel and Schubert

* On leave from: Information Science and Scientometrics Research Unit (ISSRU), Library of the Hungarian
Academy of Sciences, P.O.Box 1002, H-1245 Budapest, Hungary.
(1992, 1995) and Glänzel and Schoepflin (1994, 1995). This model assumed a non-homogeneous birth-process and was especially developed for application to bibliometric citation processes.

Of course, both Burrell’s and Glänzel’s approach were closely connected with the obsolescence phenomenon of scientific information. Thus the latter model was rather applied to explain and measure ageing of scientific literature (Glänzel and Schoepflin, 1994, 1995) than to truly predict future citation impact of publications. One reason was the relatively small data sets and the short observation period of ten years underlying the empirical studies. The second problem lied in insufficient statistical reliability of the 3 estimated parameters (α, β and N) of the assumed stochastic process if the initial observation period is short. In the following, we will show that sufficiently reliable medium-term predictions based on relatively short observation periods and using more empirical techniques are nevertheless possible.

2. The non-homogeneous citation process by Glänzel, Schubert and Schoepflin

The genesis of the assumed stochastic citation process has exhaustively explained by Glänzel and Schubert (1995) and Glänzel and Schoepflin (1994). However, for the sake of the intelligibility of the following treatment, it is indispensable to briefly summarise basic results and formulae.

Let $X_y(t)$ denote the (random) number of citations received by a paper published in a given year $y$. To simplify matters we put $y = 0$. This assumption does not result in any loss of generality. Then $X(t) = X_0(t)$ is the citation process of an arbitrary paper published in the given year ($y = 0$), and $P(X(t) = i)$ is the probability that it has received exactly $i$ citations in the interval $(0, t)$. Citation dynamics is basically reflected by the distribution $P(X(t) = i), t \in [0, \infty)$ of the process. The postulates of the Glänzel–Schubert–Schoepflin model result in the following probability distribution.

$$P(X(t) = k) = \left( N + k - 1 \right)^{k-1} \binom{N + k}{k} \left( 1 - e^{-rt} \right)^k; k \geq 0$$

with $r(t) = \beta(1-e^{-\alpha t})$ for some $N, \alpha, \beta > 0$, that is, the distribution of citations over papers is negative binomial at any time. In particular, for the transition probabilities we have in any time period $t > s \geq 0$