Price index and its relation to the mean and median reference age.

THE PRICE INDEX AND ITS RELATION TO THE MEAN AND MEDIAN REFERENCE AGE

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium(*)
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ABSTRACT

The paper consists of two parts. In the first part we assume the simple decreasing exponential model for aging. In this case we prove that the Price Index (the fraction of the references that are not older than a certain age) is a function of the mean reference age and also a function of the median reference age. Both functions are convexly decreasing, are 1 in 0 and tend to zero for the argument tending to infinity.

In the second part, the more realistic lognormal aging model is used. We now show that the Price Index is not a pure function of the mean or median reference age but a well defined relation in the form of a typical cloud of points. This cloud (as e.g.

(*) Permanent address.

Key Words : Price Index, mean reference age, median reference age.

Acknowledgement : The author is indebted to prof. Dr. R. Rousseau for stimulating discussions on this topic and to Prof. Dr. P. Janssen for information on Hoeffding's theorem.

discussed in a recent paper of Glänzel and Schoepflin) is explained using results from probabiliy theory and statistics. New data (about reference ages in JASIS) are produced that confirm the theoretical findings.

I. INTRODUCTION

Price, in his classic paper Price (1970) defines the so-called "Price Index" as "the proportion of the references that are to the last five years of literature". It is hereby unimportant whether or not we use the references of one article or the references of all the articles in a journal or the references of all the articles in all journals in a certain discipline. Of course, the result may be different whether we consider a discipline to be a set of articles or to be a set of journals (see for this a recent paper by Egghe and Rousseau (1995)), but in this paper we assume that we have a fixed set of references.

Glänzel and Schoepflin (1995) deal with the Price Index as the proportion of references that are not older than 2 years.

Therefore we can define, more generally the Price Index PI_d as the fraction of references that are 0,1,2,..., d years old.

In Glänzel and Schoepflin (1995) a graph of PI_2 (× 100) versus the mean reference age is produced. With permission of the authors, this graph is presented again (figure 1).

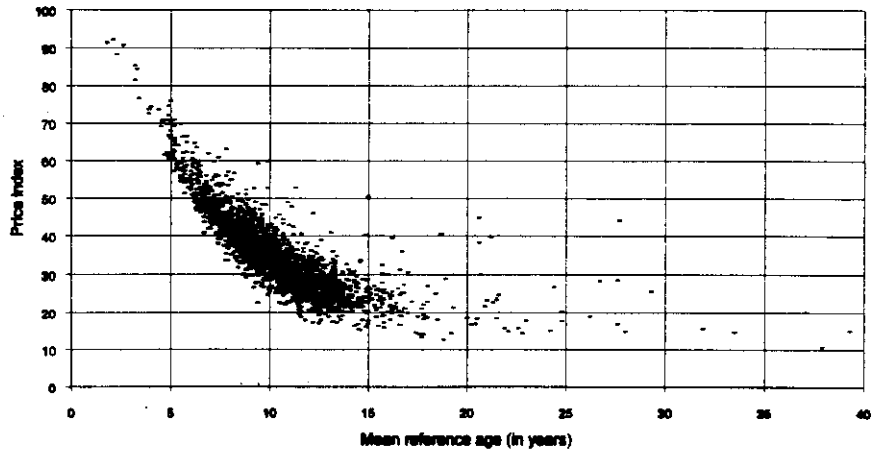

Figure 1 : Plot of "Price Index" versus mean reference age

This is a remarkable graph. First of all, the overall tendency is convexly decreasing although the relation is not a pure function : indeed the middle part contains an apparently thick cloud of points, which seems to be structural and it smoothly disappears in the beginning and the end of the graph. Furthermore it seems to be clear that $PI_2 \approx 1$ for the mean reference age going to zero (the graph says 100 %, being a fraction 1) and that PI_2 goes to zero for the mean reference age going to infinity. How can this be explained?

A good scientific method consists of starting with known results on reference age distribution and then try to explain the above graph. The simplest basic law on reference age distribution is the exponential law $c(t) = c\ a^t$ with c(t) = the distribution of the number of references that are t years old and where $0 < a < 1$. Assuming this law gives a first explanation of the graph in figure 1 : we prove that PI_d (any $d \in \mathbb{N}$) is a convexly decreasing function of the mean reference age MA which is 1 in 0 en goes to 0 if the mean reference age goes to ∞. The same conclusion can be drawn when the relation of PI_d with the median reference age (the so-called half-life) is studied. This explains the shape of the graph but certainly not the thick cloud in the middle part.

In the next section we then use the more realistic form of c(t), namely a lognormal distribution. It has been proved in Egghe and Rao (1992a) that c(t) indeed can be best modelled via a lognormal distribution and there was also given a theoretical explanation for this. The theory was confirmed by Matricciani (1991) on the age distribution of references in engineering papers.

Now it can be shown that PI_d is not a function of the mean reference age (MA) anymore but a mathematical relation (a cloud). The shape of this cloud is explained to be convexly decreasing (as is the case with the corresponding function in case we use the exponential model).

We have executed the same study for the median reference age MD, again using the lognormal distribution for c(t). Again we found that PI_d is not a function of the median reference age anymore but that the cloud behaves as in the MA case.

The "median" analogue of figure 1 was not readily available. Therefore, we collected some new data that are presented in the last section : we studied the reference lists in JASIS from 1986 until the current issue in 1995 (issue 7) : each article then gives a mean and median reference age, and several possible Price Indexes PI_d. We present PI_2, PI_4 and PI_5 versus the mean and the median reference age. These graphs confirms the theoretical findings of the previous section.

II. THE CASE THAT THE REFERENCE AGE DISTRIBUTION IS EXPONENTIAL

In this section we suppose that we have a set of references with age distribution given by

$$c(t) = c.a^t \qquad (1)$$

$t = 0,1,2,3,...$ and $0 < a < 1$. Since c(t) is a distribution one has

$$\sum_{t=0}^{\infty} c\, a^t = 1$$

Hence

$$c = a - 1 \qquad (2)$$

This is a basic function, not only in mathematics but also in informetrics : it is the basic function for aging ($0 < a < 1$) and for growth ($a > 1$) from which the more realistic models are a deviation (lognormal distribution in the case of aging, an S-shaped like function such as the logistic curve or Gompertz distribution in the case of growth - see Egghe and Rao (1992a,b)). In addition to this, this paper will prove that from (1), basic results on other parameters (such as the Price Index) can be derived.

For $d = 1,2,3,...$ we define the Price Index

$$PI_d = \frac{\#\text{ references that are } 0, 1, 2, \ldots, d \text{ years old}}{\text{total } \#\text{ references}}$$

In Price (1970), PI_4 or PI_5 is used (this is not clear from Price's text, although Wouters and Leydesdorff (1994) use PI_5 in reference to Price (1970)). In Glänzel and Schoepflin (1995), PI_2 is used.. The specific index d is not so important : our results will be valid for every $d \in \mathbf{N}$.

We will study PI_d as possible function of MA, the mean reference age and of MD, the median reference age. In both cases we will use the following result.

Theorem II.1 :

In case (1) is valid, we have

$$PI_d = 1 - a^{d+1} \tag{3}$$

Proof :

By definition

$$PI_d = (1-a)\,[1 + a + \ldots + a^d]$$

(since $\sum_{t=0}^{\infty} c\, a^t = 1$)

and hence

$$PI_d = 1 - a^{d+1}$$ □

We now start with the MA-dependency.

II.1. PI_d as function of MA

By definition,

$$MA = \sum_{t=0}^{\infty} t\,(1-a)\,a^t$$

$$= (1-a)(a + 2a^2 + 3a^3 + \ldots)$$

$$= (1-a)(a + a^2 + a^3 + \ldots$$

$$+ a^2 + a^3 + \ldots$$

$$+ a^3 + \ldots$$

$$+ \ldots)$$

$$(1-a)\,[a(1 + a + a^2 + \ldots) + a^2(1 + a + a^2 + \ldots) + a^3(1 + a + a^2 + \ldots) + \ldots$$

$$= (1-a)(1+a+a^2+\ldots)(a+a^2+a^3+\ldots)$$

$$= (1-a)\ \frac{a}{(1-a)^2}$$

hence

$$MA = \frac{a}{1-a} \tag{4}$$

Consequently,

$$a = \frac{MA}{1+MA} \tag{5}$$

(3) and (5) yield the following theorem.

Theorem II.2 :

$$PI_d = 1 - \left(\frac{MA}{1+MA}\right)^{d+1} \tag{6}$$

This function is convexly (*) decreasing,

$$\lim_{MA \to 0} PI_d = 1, \qquad \lim_{MA \to \infty} PI_d = 0$$

(*) Convexity is guaranteed if MA > d/2 which is always the case for reasonable d.

Proof :

Formula (6) follows readily from (3) and (5). Furthermore, taking derivatives w.r.t. MA yields

$$PI_d' = -\frac{(d+1)(MA)^d}{(1+MA)^{d+2}} < 0$$

(always) and

$$PI_d'' = (d+1)(MA)^{d-1} \frac{2\,MA - d}{(1+MA)^{d+3}} > 0$$

if MA > d/2. We can assume that this is the case : in the Glänzel and Schoepflin case MA is required to be larger than 1, in the Price case MA must be larger than 2 or 2.4, which will be true in almost all cases. If MA < d/2, then we loose the convexity property. In any case, it is trivial that

$$\lim_{MA \to 0} PI_d = 1 \tag{7}$$

$$\lim_{MA \to \infty} PI_d = 0 \tag{8}$$

□

Basically, this explains the shape of the Glänzel-Shoepflin graph. The same results follow if we replace MA bij MD, the median references age (being also the half-life).

II.2. PI_d as function of MD

Let us denote by C(t) the cumulative function.

$$C(t) = \sum_{t'=t}^{\infty} c(t')$$

$$C(t) = \sum_{t'=t}^{\infty} (1-a)\,a^{t'}$$

$$C(t) = (1-a)\,a^{t}\,(1+a+a^2+\ldots)$$

$$C(t) = a^{t}$$

By definition, MD is this value of t for which

$$C(MD) = \frac{1}{2} C(0)$$

This gives :

$$a^{MD} = \frac{1}{2}$$

hence

$$MD = \frac{\ln (0.5)}{\ln a} \qquad (9)$$

This formula also appears in Egghe and Rousseau (1990), p.270.
From this

$$a = (0.5)^{\frac{1}{MD}} \qquad (10)$$

(10) and (3) yield the following theorem.

Theorem II.3 :

$$PI_d = 1 - (0.5)^{\frac{d+1}{MD}} \qquad (11)$$

This function is convexly (*) decreasing,

$$\lim_{MD \to 0^{>}} PI_d = 1, \qquad \lim_{MD \to \infty} PI_d = 0$$

(*) Convexity is guaranteed if MD > ((ln 2)/2) (d+1) which is always the case for reasonable d.

Proof :
Formula (11) follows readily. Deriving w.r.t. MD gives

$$PI_d' = (0.5)^{\frac{d+1}{MD}} \ln (0.5) \frac{d+1}{(MD)^2} < 0$$

$$PI_d'' = -\ln(0.5)(0.5)^{\frac{d+1}{MD}} \frac{d+1}{(MD)^3}\left(2 + \frac{d+1}{MD}\ln(0.5)\right) > 0$$

if

$$MD > \frac{\ln 2}{2}(d+1) \approx 0.3466\ (d+1)$$

This is clearly satisfied in most cases for d = 2 is the condition MD > 1.04, for d = 4 we have MD > 1.73 and for d = 5, MD > 2.08, what is true in almost all cases. If this condition is not satisfied we only loose the convex shape of the curve. Further it is clear that

$$\lim_{MD \to 0} PI_d = 1 \qquad (12)$$

and

$$\lim_{MD \to \infty} PI_d = 0 \qquad (13)$$

□

So, this section fully explained the general shape of the functions MA → PI_d and MD → PI_d if the reference age distribution is given by (1). The next section deals with the lognormal distribution, the natural distribution describing reference ages (Egghe and Rao (1992a) and Matricciani (1991)).

III. THE GENERAL CASE : THE CASE THAT THE REFERENCE AGE DISTRIBUTION IS LOGNORMAL

III.1. Introduction

It is well-known that the reference age follows a lognormal distribution. This means that ln t is normally distributed. We hence have (cf. Goldberg (1984), p.404) :

$$c(t) = \frac{1}{t\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right] \tag{14}$$

where μ and σ is the mean resp. the standard deviation of ln t.

Since PI_d is the fraction of references that are not older than d years we have

$$PI_d = \int_0^d \frac{1}{t\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right] dt \tag{15}$$

We can simplify this expression as follows : substitute x = ln t and after this

$$Z = \frac{x - \mu}{\sigma}$$

We then have

$$PI_d = \int_{-\infty}^{\frac{\ln d - \mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}\, dZ \tag{16}$$

Consequently,

$$PI_d = F\left(\frac{\ln d - \mu}{\sigma}\right) \tag{17}$$

where F denotes the cumulative standard normal distribution function.

Next we express the mean reference age MA and the median reference age MD. In Goldberg (1984), p.405 one finds :

$$MA = e^{\mu + \frac{\sigma^2}{2}} \tag{18}$$

For MD we have, by definition

$$P(0 \leq t \leq MD) = 0.5$$

Hence

$$P(-\infty \leq \ln t \leq \ln MD) = 0.5$$

and

$$P\left(-\infty \leq \frac{\ln t - \mu}{\sigma} \leq \frac{\ln MD - \mu}{\sigma}\right) = 0.5$$

Since $\frac{\ln t - \mu}{\sigma}$ is standard normally distributed we hence have

$$\frac{\ln MD - \mu}{\sigma} = 0$$

Consequently

$$MD = e^{\mu} \tag{19}$$

It is trivial to see that MD < MA.

Formulae (18) and (19) show that both functions $(\mu,\sigma^2) \rightarrow MA$ and $(\mu,\sigma^2) \rightarrow MD$ are not injective. This is the more clear when we express PI_d in terms of MA resp. MD. Formulae (15) and (18) imply

$$PI_d = \int_0^d \frac{1}{t\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \ln MA + \frac{\sigma^2}{2}}{\sigma}\right)^2\right] dt$$

$$PI_d = F\left(\frac{\ln d - \ln MA + \frac{\sigma^2}{2}}{\sigma}\right) \tag{20}$$

Here σ still can vary, yielding different values of PI_d for one value of MA.
For MD we have :

$$PI_d = \int_0^d \frac{1}{t\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \ln MD}{\sigma}\right)^2\right] dt$$

$$PI_d = F\left(\frac{\ln d - \ln MD}{\sigma}\right) \qquad (21)$$

yielding the same conclusion as above.

So PI_d is not a function of MA nor of MD. Still we want to explain the cloud of points in figure 1 (w.r.t. the variable MA) and the analogous graph w.r.t. the variable MD.

In the next section we will explain the general **shape** of the graph in figure 1 (and the same for the variable MD).

III.2. Shape of cloud of points MA → PI_d

Since the relation MA → PI_d is not a function, we cannot study the MA-dependency in formula (20) just like that. What we can do is study the set of functions (that constitute the whole graph of MA → PI_d) as in (20) but with

$$\sigma = \text{constant} \tag{22}$$

This gives trajectories in this cloud of points that can be studied as follows.

Using formula (20) and using the fact that

$$F'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{23}$$

we have, taking the derivative w.r.t. MA and using (22)

$$PI_d' = \frac{1}{\sqrt{2\pi}} \exp\left[- \frac{\left(\frac{\ln d - \ln MA + \frac{\sigma^2}{2}}{\sigma} \right)^2}{2} \right] \left(- \frac{1}{\sigma . MA} \right)$$

< 0, always

$$PI_d'' = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma (MA)^2} \exp\left[- \frac{\left(\frac{\ln d - \ln MA + \frac{\sigma^2}{2}}{\sigma} \right)^2}{2} \right] \cdot \left[1 - \frac{\ln d - \ln (MA) + \frac{\sigma^2}{2}}{\sigma^2} \right]$$

> 0

if and only if

$$MA > \frac{d}{e^{\frac{\sigma^2}{2}}} \tag{24}$$

which is always satisfied if MA ≥ d. This is clearly true in practical cases and for the common values of d (e.g. d ≤ 5). It is certainly so for d = 2 in the Glänzel and Schoepflin paper.

In fact one could also argue that, by the very nature of the Price Index (knowing the fraction of the "younger" references) it does not make much sense to study PI_d for d > MA.
Furthermore

$$\lim_{\substack{MA\to 0 \\ >}} PI_d = \lim_{\mu\to-\infty} PI_d = F(+\infty) = 1$$

since (by (22)) MA → 0 if and only if $\mu \to -\infty$. Also MA → $+\infty$ if and only if $\mu \to +\infty$. Now

$$\lim_{MA\to\infty} PI_d = F(-\infty) = 0$$

So, on the trajectories σ = C in the cloud of points as in figure 1 we found convexly decreasing functions which are 1 in 0 and tend to 0 for MA → $+\infty$: the same conclusions as in section I. This could be expected since the lognormal distribution is only a "perturbation for small t" of the exponential distribution. Yet, of course, the results of section I are not immediately valid, without extra proof (as given here), in this section.

III.3. Shape of cloud of points MD → PI_d

It is now easy to do the same for the variable MD. Formula (21) yields, taking the derivative w.r.t. MD (using (22)) :

$$PI'_d = \frac{1}{\sqrt{2\pi}} \exp\left[- \frac{\left(\frac{\ln d - \ln MD}{\sigma}\right)^2}{2} \right] \left(- \frac{1}{\sigma . MD} \right)$$

< 0, always

$$PI''_d = \frac{1}{(MD)^2} \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left[- \frac{\left(\frac{\ln d - \ln MD}{\sigma}\right)^2}{2} \right] . \left[1 - \frac{\ln d - \ln MD}{\sigma^2} \right]$$

> 0

if and only if

$$MD > \frac{d}{e^{\sigma^2}} \qquad (25)$$

Note that this condition is the same as (24). For this reason, again MA ≥ d suffies (and see the remarks made below formula (24)).

Again MD → 0 if and only if $\mu \to -\infty$ and MD → $+\infty$ if and only if $\mu \to +\infty$, yielding

$$\lim_{\substack{MD \to 0 \\ >}} PI_d = F(+\infty) = 1$$

and

$$\lim_{MD \to \infty} PI_d = F(-\infty) = 0$$

So the graph of MD → PI_d has on its trajectories $\sigma = C$ the same shape as the one of MA → PI_d. These findings will be confirmed by the practical data of section IV.

Note : We hope that the results of section I and the ones given above are enough explanation for the cloud of points as in figure 1. Of course, theoretically it can be that the studied trajectories do not have the same shape as the overall cloud, but figure

1 rejects this possibility. It is hence clear that the graph of figure 1 is "composed" of different trajectories σ = C, where C varies. This is depicted in figure 2.

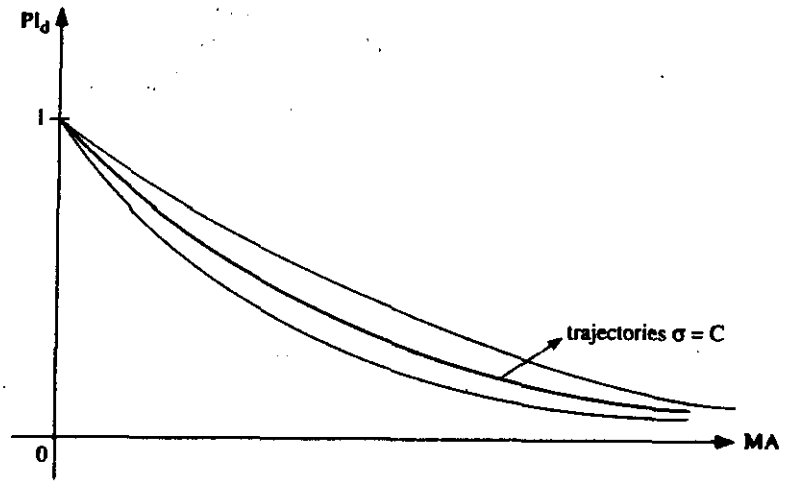

Figure 2 : Decomposition of figure 1 into trajectories σ = C

The same can be said about the relation MD $\rightarrow$ PI_d.

IV. NEW EXPRIMENTAL DATA ON MEAN AND MEDIAN REFERENCE AGE AND ON THE PRICE INDEX

We have examined the reference lists of every article in the journal JASIS in the period : 1986 until the present (issue 7 of 1995). We excluded letters to the editor or book reviews and also the review articles in the "Perspective" issues. The reason is that we wanted to collect homogeneous material so that the age distribution of every reference list can be viewn as a sample of a general lognormal distribution. We took JASIS since most of the articles in it have substantive reference lists, contrary to e.g. Scientometrics where regularly articles with very short reference lists appear.

For each article we calculated

- the total number of references
- the number of references to the years j, j-1, j-2, j-3, j-4 and j-5, where j denotes the year of publication of the article
- the mean reference age
- the median reference age.

In total we studied 367 articles (we also left out two historical articles : they could cause scaling problems on the mA - PI_d and MD - PI_d graphs; furthermore we wanted articles as homogeneous as possible as explained above).

The mean reference age (say m) was transformed into

$$MA = j - m$$

so that time has a fixed origin and so that reference ages could be compared and put in the same graph. The same was done for MD.

By taking the number of references to the years j, j-1 and j-2, adding these three numbers and dividing the result by the total number of references, we obtain, for each article, a PI_2 value. Going two years further we obtain PI_4 and adding one more year (j-5) yields PI_5 in the same way. These PI-values where then put into a

graph versus the MA- and MD-values. We hence obtained six graphs - see figures 5 -10.

All six graphs show a convexly decreasing cloud of points. The time period used to calculate PI_2 apparently is too small to yield values above 0.8. These values (up to 1) are obtained in the PI_4- and PI_5-case.

The thickness in the middle part of the graphs (with MA as abscis) is apparent. It is also clear that the MD-graphs show more variability of PI-values for small MD-values than is the case for small MA-values. For high MD-values this cannot be concluded, due to the scarceness of data.

These graphs, consequently, confirm our theoretical findings.

The graphs were obtained using the statistical package STATGRAPHICS 6.0.

Figure 3 : Plot of PI_2 versus MA

Figure 4 : Plot of PI_2 versus MD

Figure 5 : Plot of PI_4 versus MA

Figure 6 : Plot of PI_4 versus MD

Figure 7 : Plot of PI_5 versus MA

Figure 8 : Plot of PI_5 versus MD

The values of PI are multiplied by 100.

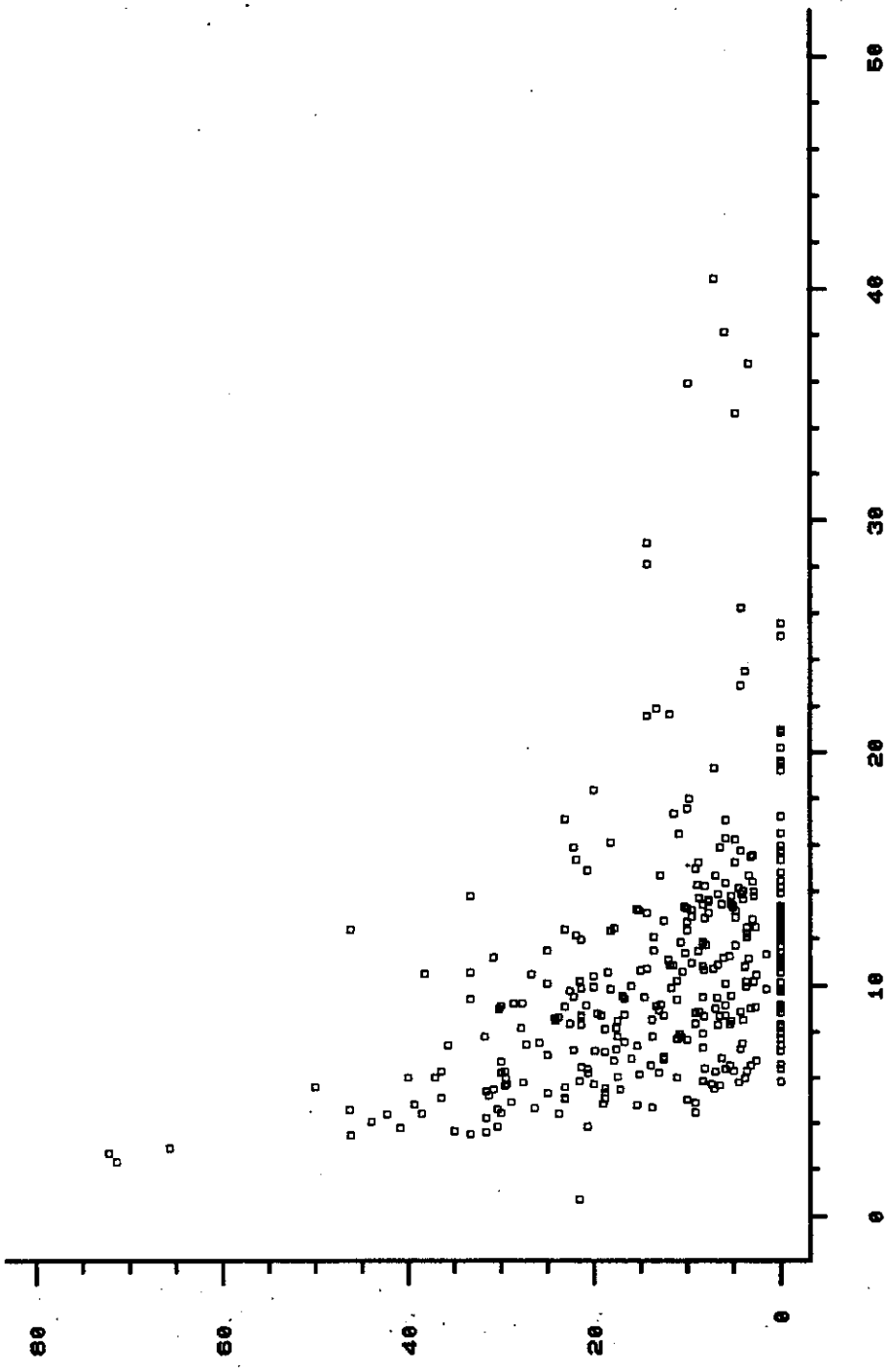

Fig. 3

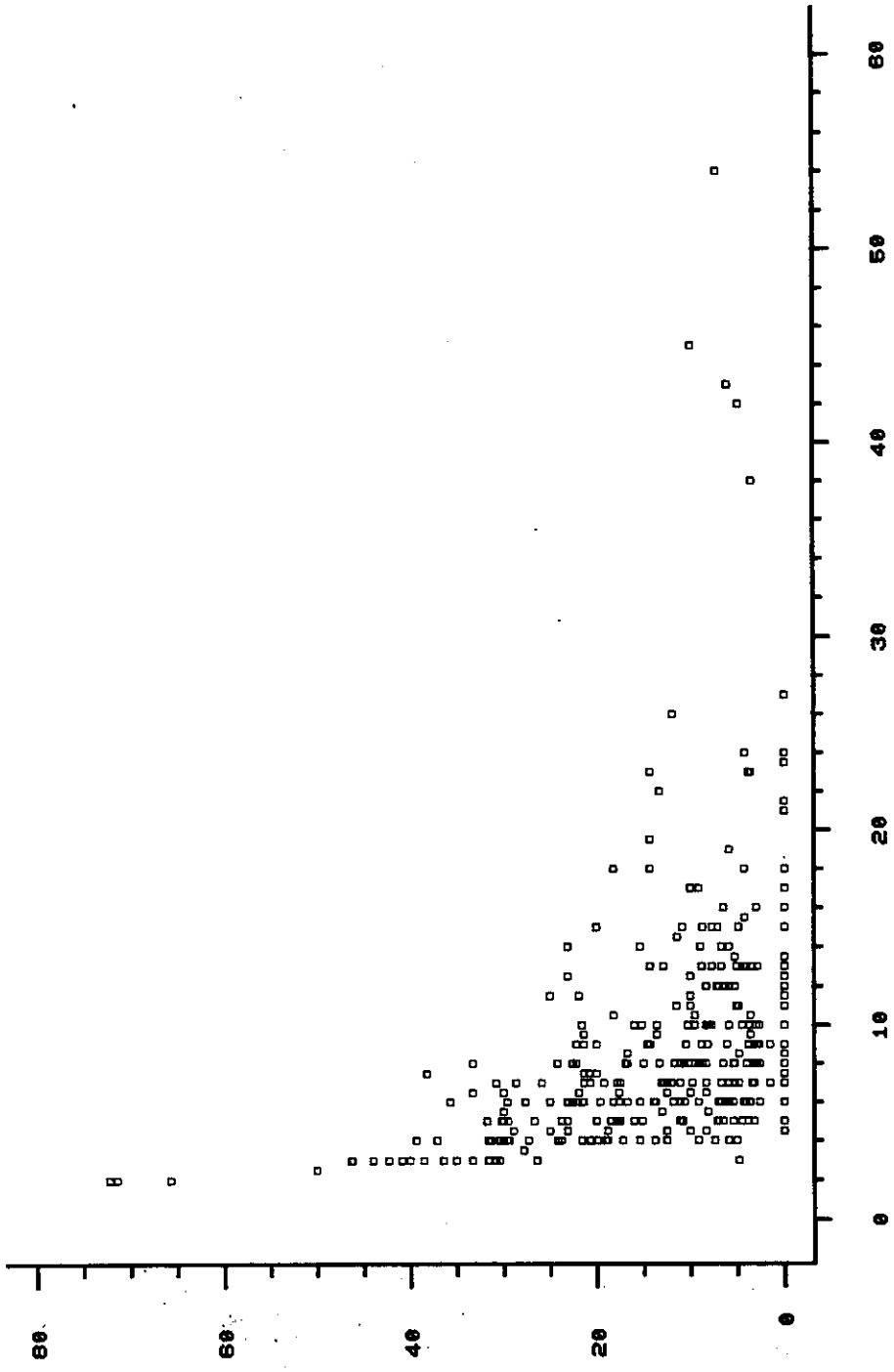

Fig. 4

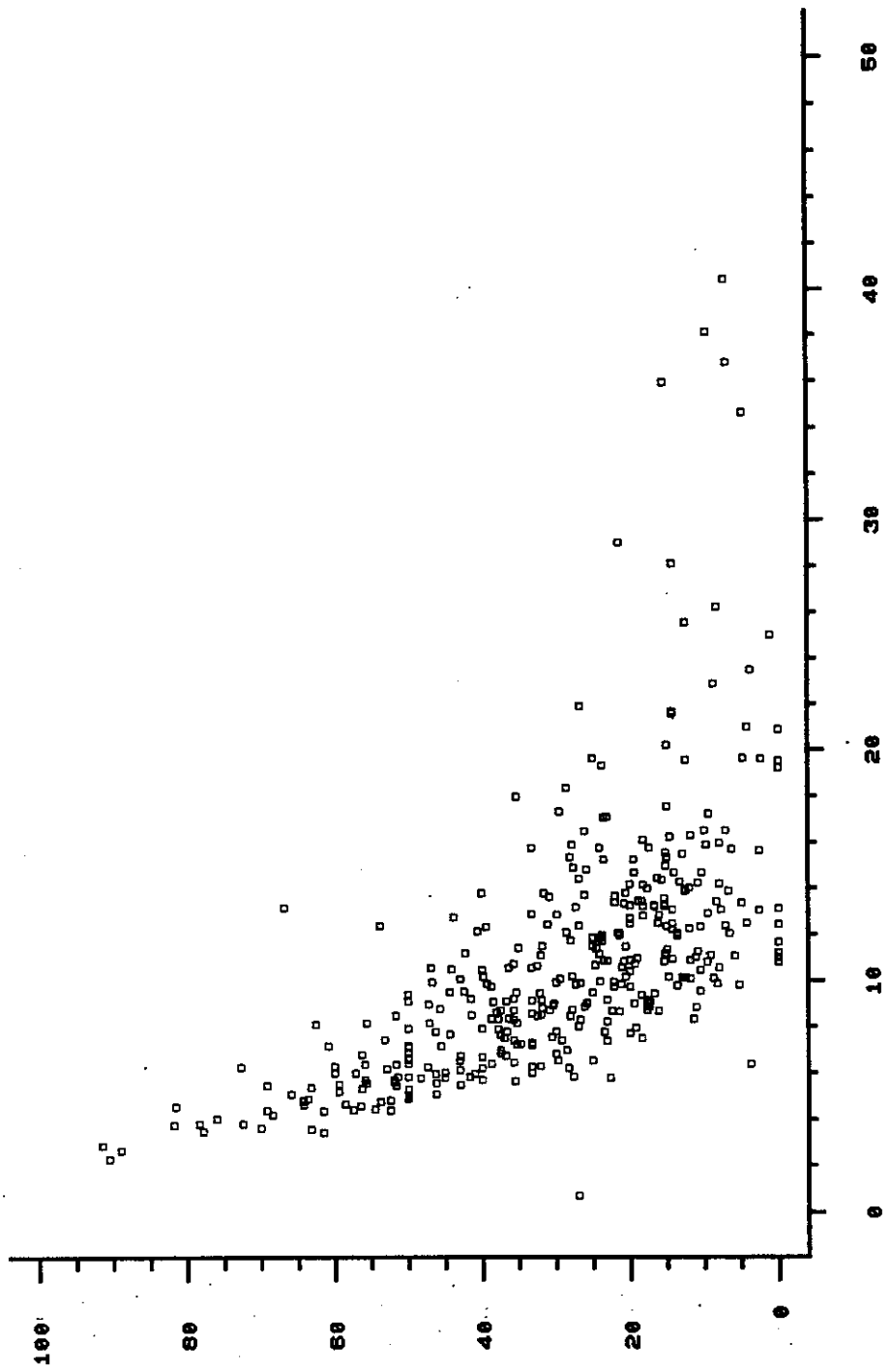

Fig. 5

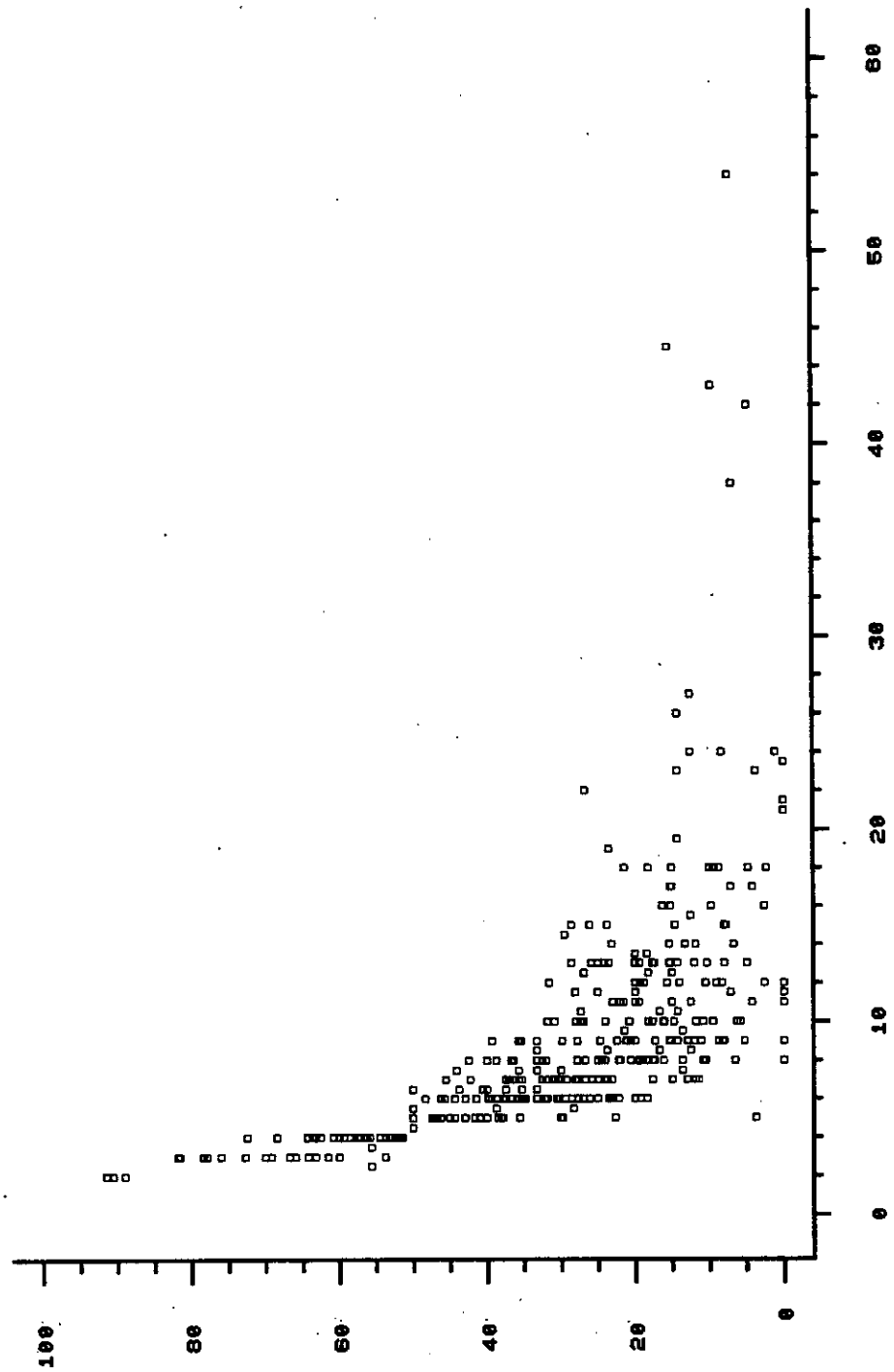

Fig. 6

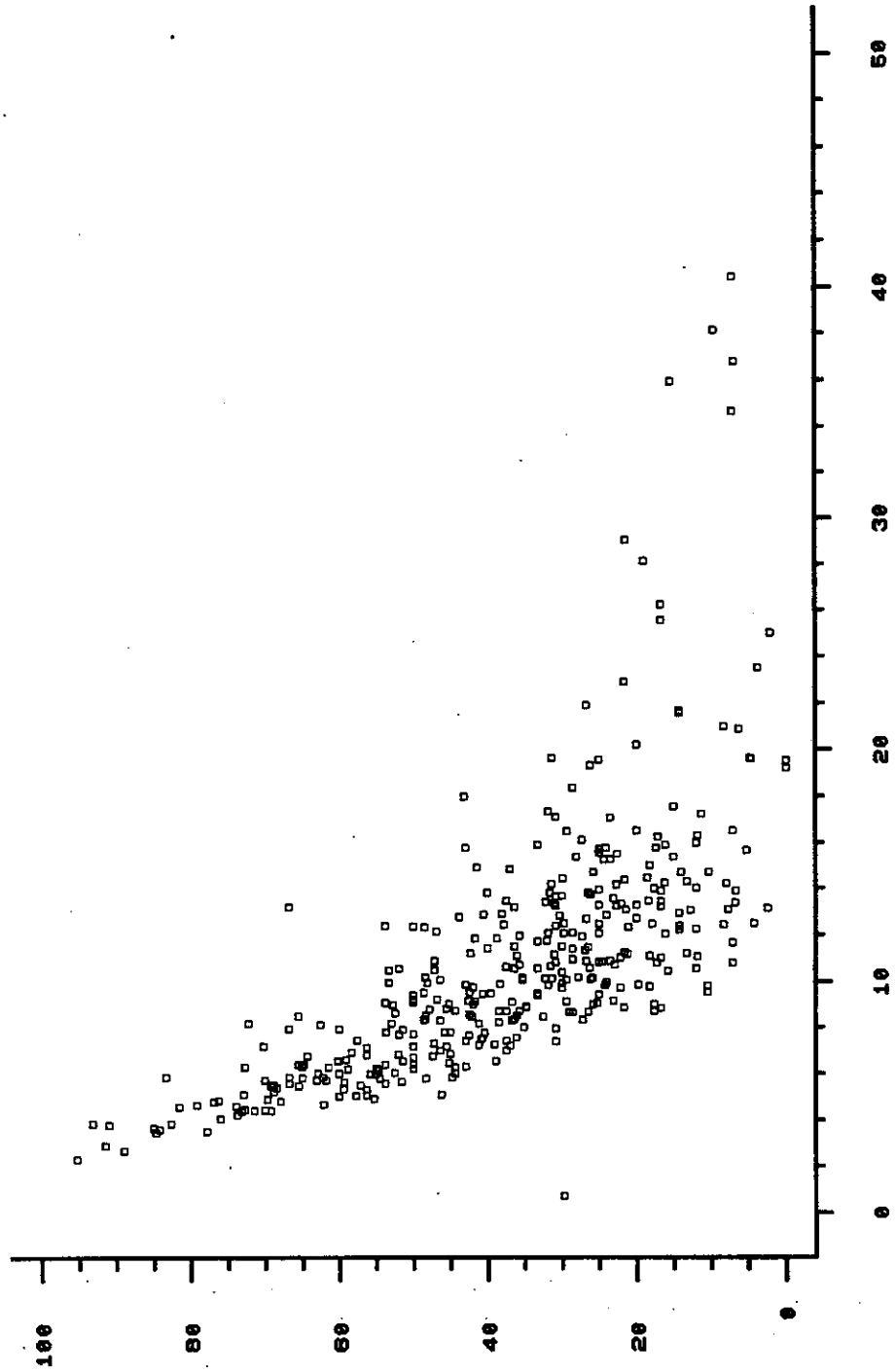

Fig. 7

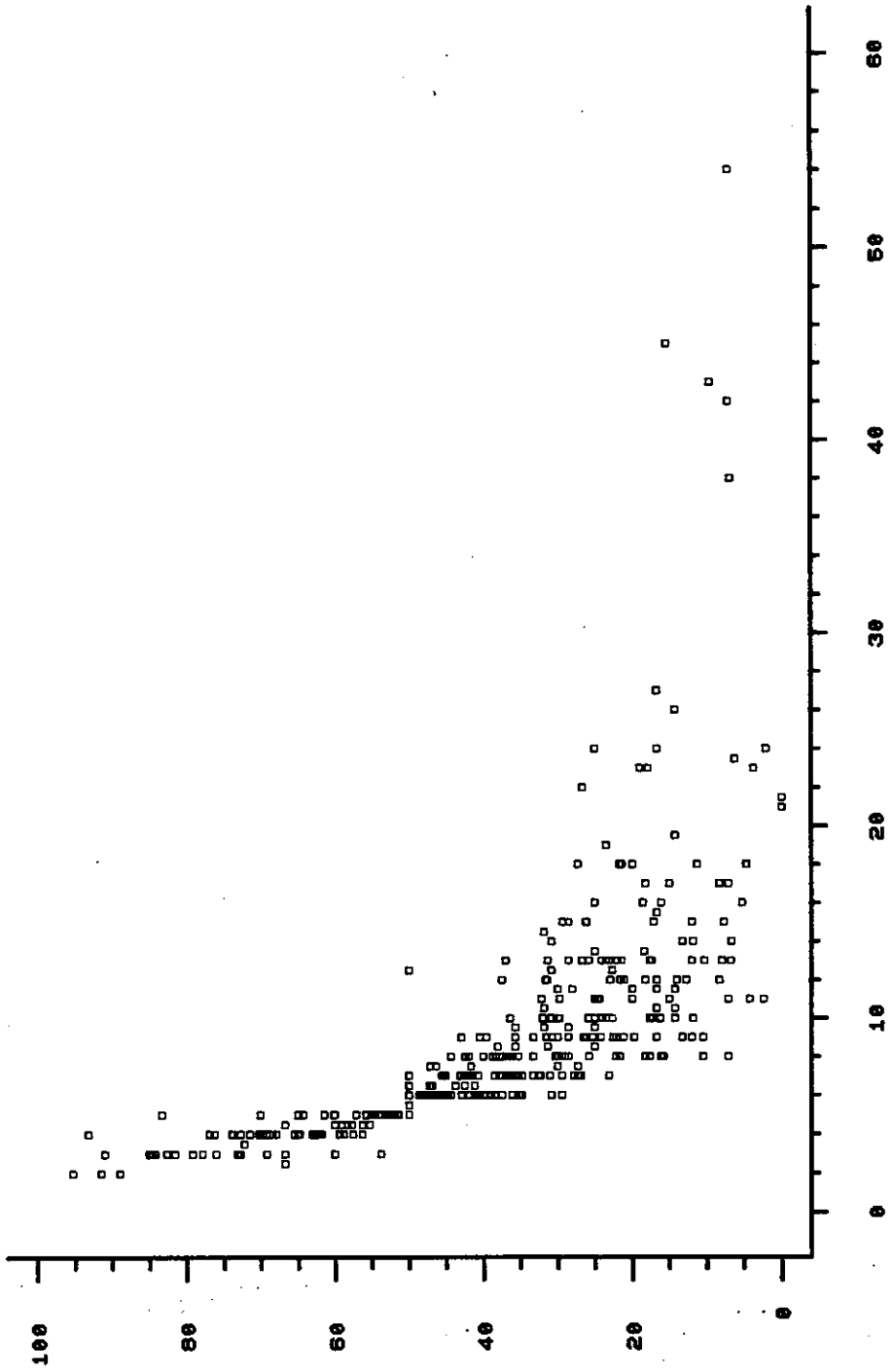

Fig. 8

REFERENCES

L. Egghe and I.K.R. Rao (1992a). Citation age data and the obsolescence function : fits and explanations. Information Processing and Management, 28(2), 201-217.

L. Egghe and I.K.R. Rao (1992b). Classification of growth models based on growth rates and its applications. Scientometrics, 25(1), 5-46.

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

L. Egghe and R. Rousseau (1995). Average and global properties of informetric data. Preprint.

W. Glänzel and U. Schoepflin (1995). A bibliometric ageing study based on serial and non-serial reference literature in the sciences. Proceedings of the fifth biennial Conference of the International Society for Scientometrics and Informetrics, River Forest, Ill., USA (June 7-10, 1995), Learned Information, Medford, NJ, USA, 177-185.

M.A. Goldberg (1984). An Introduction to Probability Theory with Statistical Applications. Plenum Press, New York.

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

D. De Solla Price (1970). Citation measures of hard science, soft science, technology and nonscience. In : C.E. Nelson, D.K. Pollack (eds.). Communication among Scientists and Engineers, Heath, Lexington, MA, USA, 3-22.

P. Wouters and L. Leydesdorff (1994). Has Price's dream come true : is scientometrics a hard science? Scientometrics, 31(2), 193-222.