

A noninformetric analysis of the relationship between citation age and journal productivity

Peer-reviewed author version

EGGHE, Leo (2001) A noninformetric analysis of the relationship between citation age and journal productivity. In: Journal of the American Society for Information Science and Technology, 52(5). p. 371-377.

DOI: 10.1002/1532-2890(2001)9999:9999<::AID-ASI1079>3.0.CO;2-L

Handle: <http://hdl.handle.net/1942/777>

A NON-INFORMETRIC EXPLANATION OF THE RELATION BETWEEN THE JOURNAL MEAN CITATION AGE AND ITS NUMBER OF ARTICLES

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium¹

and

UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ABSTRACT

A problem, raised by Wallace (JASIS, 37, 136-145, 1986), on the relation between the journal's median citation age and its number of articles is studied. Leaving open the problem as such, we give a statistical explanation of this relationship, when replacing "median" by "mean" in Wallace's problem.

The cloud of points, found by Wallace, is explained in this sense that the points are scattered over the area in first quadrant, limited by a curve of the form

$$y = \frac{E}{x^2} + 1 ,$$

¹ Permanent address.

Keywords and phrases : mean citation age, journal productivity, Central Limit Theorem.

where E is a constant. This curve is obtained by using the Central Limit Theorem in statistics and, hence, has no intrinsic informetric foundation.

The paper closes with some reflections on explanations of regularities in informetrics, based on statistical, probabilistic or informetric results, or on a combination thereof.

I. Introduction

Detecting and explaining regularities (functions, laws, distributions, formulae,...) are the most fundamental subjects of scientific research. Based on "raw" data one usually constructs graphs of "clouds of points" which show a certain regularity. If this regularity resembles a graph or a curve, one often tries to fit such a curve to this cloud of points via - e.g. - (nonlinear) regression. Many authors then stop their investigations assuming they have given an "explanation" of the observed facts. This is only partially true - or, I would prefer to say : this is partially wrong. The classical "joke" is fitting the first 10 or 20 points on the graph $y=\sqrt{x}$ for $x \in \mathbb{N}$, the set of positive entire numbers. It is an easy exercise to show that a linear fit works well with a highly significant value of the correlation coefficient of Pearson. Authors stopping here would then pretend to have shown that $y=\sqrt{x}$ ($x=1,2,\dots, 20$, say) is a linear relationship !

Is it wrong to apply such fitting techniques ? Not at all ! They provide first basic information on the shape of a cloud of points, although several different fittings (with different regularities) are possible (see the above example !). This is - however - only the first part in the scientific work. The next part is to start a kind of rationale, leading to - preferably - one of the previously obtained regularities (functions, distributions). A rationale is only possible if one knows where to depart from - i.e. what are the laws (regularities or other tools) that are (or can be) presupposed. In other words, given previously established results (or, at least, given some acceptable axioms), try to deduce (by reasoning) the (or some) observed regularities. This is then as in pure mathematics, where

one deduces from axioms (the notion of "acceptable" is not even existing here) results which can then again be used to deduce further results such as regularities.

A very simple example of this methodology is given in Egghe and Rao (1992a) where one tries to explain aging curves : the basic aging curve, being the exponential decay, is generalized so that the initial fast increase of use (and then followed by a slow decrease as in the case of an exponential decay) is also explained. In this reference one uses the lognormal distribution, which is fully explained in probability theory (the argument is even repeated in Egghe and Rao (1992a)). In a similar way, growth can be studied - see Egghe and Rao (1992b).

Another "basic" example is the explanation of the classical informetric laws as e.g. the ones of Lotka (1926), Bradford (1934), Zipf (1949), Mandelbrot (1954, 1977) and so on. For some of these explanations we refer the reader to Egghe and Rousseau (1990), where several explanations, given by different authors, are reproduced (explanations given by De Solla Price (1976), Mandelbrot (1954, 1977), Bookstein (1977, 1990a,b)). These informetric laws all deal with the relationship between sources and items (items being produced by sources). Classical examples are : articles "produced" by journals or authors, citations "produced" by articles (here citations can be given or being received), and so on. In linguistics, one can talk about words (type) and their uses in texts (token). So these laws are also called type-token relations (see e.g. Herdan (1960)). In informetrics they are (usually) called source-item relations, but essentially they have the same meaning or interpretation.

A further step can be to use these above results and combine them in a way as to explain further regularities. Examples (and reviews) of this are given in the last section.

But before we do this, the next section is devoted to an apparent informetric problem raised by Wallace (1986). Wallace writes literally (p. 137) : *"For a given subject literature, the median citation ages of the journals contributing to that literature will vary inversely with the productivity of thoses journals, where productivity is measured in terms of the number of articles contributed by each journal"*. If this is true it would be a remarkable informetric

result, useable in many other applications. In any case, we are in need of an explanation. The found regularity is shown in the graph of Figure 1 (reproduced with permission).

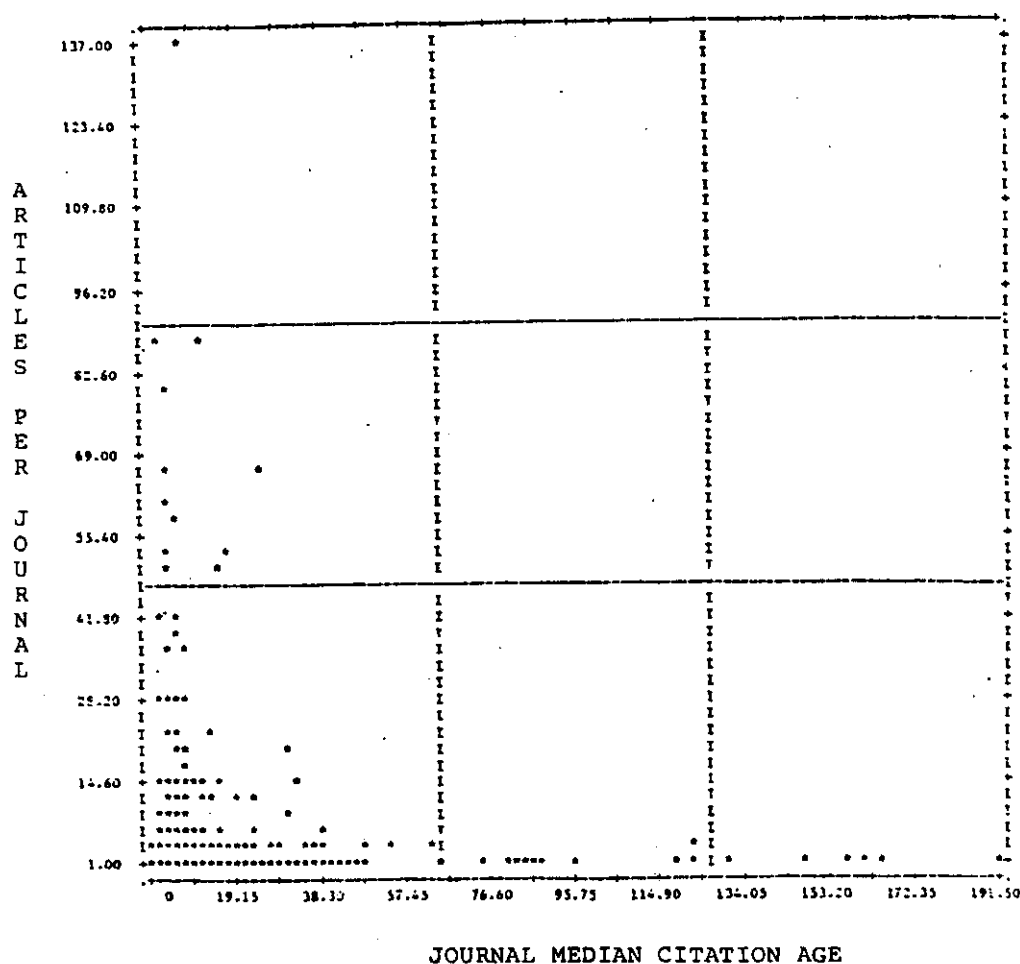


Fig.1. Journal productivity plotted against journal median citation age.

A first glance indeed indicates a decreasing relationship although it is also immediately clear that the cloud of points fills an area in the first quadrant delimited by a decreasing function. In other words, the cloud of points satisfies $y \leq f(x)$ (with $x, y \geq 0$) rather than $y = f(x)$, where f is this decreasing function. This simple remark shows already that (as is clear by visual inspection and as is also remarked in Wallace (1986)) that the smaller median citation ages

occur for highly as well as less productive journals. The hypothesis as such is hence already rejected but still one wants to know why we have a relation of the form $y \leq f(x)$, and we also want to know the form of f . In Wallace (1986) an indication is given for $f(x) = \frac{c}{x}$ ($c = \text{a constant}$) but any $f(x) = \frac{c}{x^\alpha}$ ($\alpha > 0$) or another decreasing function is possible. Since we (more or less) rejected the hypothesis, we had the impression that an informetric explanation of $y \leq f(x)$ was not possible.

In the next section we indeed show that an explanation can be given by only using the Central Limit Theorem in probability theory, hence showing that the regularity in Fig. 1 probably has no informetric basis. This is no criticism on the work of Wallace (1986) - on the contrary : as many as possible regularities must be detected, but then their explanations need to be classified into informetric or not (i.e. statistical, probabilistic, informetric or a combination thereof). This will be done at the end of the paper, where we examine a few explanations that exist in the literature.

II. The relation between journal median (mean) citation age and the number of articles in a journal.

From an informetric point of view it is hard to believe that Wallace's hypothesis is true. Although it is true that - in any discipline - the median (mean) citation ages might differ from journal to journal, it would be a strange fact that these ages are related (let alone inversely related) with journals' productivity (as measured in terms of articles per journal). Only in special cases this might be true, e.g. where the top journals in the field are the largest ones and where the articles are used so heavily so that they have the smallest median (mean) citation ages (the median (mean), measured over all articles in the journal). An opposite example can also be given and this in any discipline : the case of review journals. They usually are large (in terms of number of articles) but their median and mean citation ages usually are large too.

In order to start a rationale on this matter we will limit ourselves to the case of mean citation ages and we will fix (as in Wallace (1986)) the subject, represented by a set of journals consisting of articles, each with their (diachronous or synchronous) citation age distribution. Considering all mean citation ages of all these articles, in any journal in this set, we can calculate their average : the mean citation age of the field. Here we have considered the field as consisting of articles : this approach is called the global one and is different from the approach where the field is consisting of journals and where we take the averages over the averages calculated per journal (cf. Egghe and Rousseau (1996a,b)). We note that even a third approach is possible where we consider the field consisting of citations (the “smallest” unit) but we will not go into this since we do not need it here.

Let us now consider the journals in this set, consisting of A articles ($A \in \mathbb{N}$, fixed). Let us call this the A -subfield of the entire field. This A -subfield again can be considered as consisting of articles (as we did above with the entire field). As above, this yields the mean citation age μ_A of the A -subfield. The hypothesis of Wallace, rephrased in this terminology, is that μ_A is a decreasing function of A . To show that this hypothesis is not necessarily linked with Fig. 1 and that the regularity in Fig. 1 can be explained in a non-informetric way we will assume that μ_A is a constant (say μ) of A , i.e. that μ_A does not depend on A . We assume this for the sake of simplicity but also (and more importantly) because explaining Fig. 1 with this assumption is the most “spectacular” explanation since, in our arguments, we deny Wallace’s hypothesis but we will be able to explain the regularity of Fig. 1. In the same way we assume that the variances σ_A^2 are also A -independent : $\sigma_A^2 = \sigma^2$, for all $A \in \mathbb{N}$.

For each $A \in \mathbb{N}$ fixed and for each journal with A articles (hence belonging to the A -subfield), we have that this journal’s mean citation age is a number which is the mean of a sample of A articles. The Central Limit Theorem (CLT) then yields the fact that this mean belongs to a $100(1-\alpha)\%$ confidence interval (around μ) of the form

$$\mu \pm Z(\alpha) \frac{\sigma}{\sqrt{A-1}} , \quad (1)$$

where $Z(\alpha)$ is the abscis such that, on the graph of the Gaussian distributioin (i.e. the standard normal distribution) the tails, determined by $Z(\alpha)$ and $-Z(\alpha)$ have a total area of α . More concretely, e.g. for $\alpha=0.05$, a 95 % confidence interval is given by

$$\mu \pm 1.96 \frac{\sigma}{\sqrt{A-1}} . \quad (2)$$

In general, the values of $Z(\alpha)$ can be read from the table of the standard normal distribution, which is available in any book on statistics or probability theory.

Expression (1) contains the key to the explanation of the graph in Fig. 1. Indeed, for each fixed $A \in \mathbb{N}$ (i.e. ordinate in Fig. 1) we have that the "sample" journals with A articles have a mean citation age between the values given by (1), for $100(1-\alpha)\%$ sure. The lower A , the larger this confidence interval. From (1) it follows that the deviation from μ to the right is equal to

$$m = Z(\alpha) \frac{\sigma}{\sqrt{A-1}} \quad (3)$$

, where m is the abscissa in Fig. 1. Since A is the ordinate, we will invert (3), yielding

$$A = \left(\frac{Z(\alpha)\sigma}{m} \right)^2 + 1$$

or, more simply

$$A = \frac{E_\alpha}{m^2} + 1 , \quad (4)$$

where E_α is a constant, decreasing with α . Equation (4) is the decreasing graph at the right side in Fig. 1. The "fading away" effect, observed in Fig. 1, when going from low values of m and A to high values of m and A is given by the different values of α and corresponding probabilities $(1-\alpha)$. Of course, the left part of (1) is usually cut-off by the requirement that $m \geq 0$ and $A \geq 1$.

Note also that, from (4), we have $\lim_{m \rightarrow \infty} A = 1$, $\lim_{m \rightarrow 0} A = +\infty$, all in agreement with the graph in Fig. 1.

In short, Fig. 1 but with “median” replaced by “mean” is explained by the high variances of small samples (and conversely by the small variances of large samples).

Note : There is one “informetric” element in the graph of Fig. 1 although this is not determinant for its shape : the fact that the cloud of points is thicker for low A than for high A , follows from the law of Lotka on the number of journals with A articles, being proportional to $\frac{1}{A^\beta}$ where $\beta \geq 1$ (see e.g. Egghe and Rousseau (1990)).

We think this gives a rationale for Fig. 1 (where “median” is replaced by “mean”) and sheds light on this regularity as being an effect of statistics and probability rather than being the consequence of an informetric law or regularity. We leave open to explain the “median”-case but are convinced that the given explanation more or less shows that also the graph of the relation between number of articles and median citation age is non-informetric in nature as well.

In the next section we will review a few other regularities that have been explained so far.

III. Explanations in informetrics.

III.1 The arcs at the end of a Leimkuhler curve.

One of the simplest regularities ever found in informetrics, but which is not an informetric regularity at all, is the fact that, at the end of a Leimkuhler curve, one detects “arcs”. A Leimkuhler curve obtains when graphing the cumulative number $R(r)$ of items in the first (largest) r sources, versus $\log r$. The graph looks as in Fig. 2 and can be found e.g. in Warren and Newill (1967), Brookes (1973), Praunlich and Kroll (1978), Wilkinson (1973), Summers (1983).

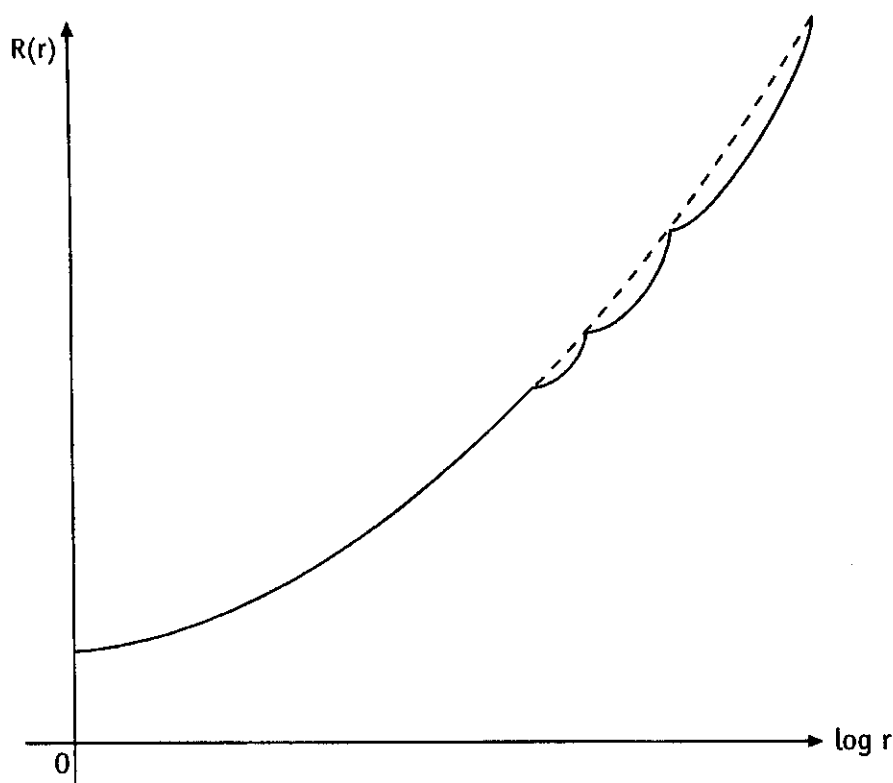


Fig. 2. A Leimkuhler curve, with arcs for large r

While the graph (without the arcs) has an equation of the form

$$R(r) = a \log (1+br) \quad (5)$$

which is certainly an informetric regularity (see Egghe (1989,1990), Egghe and Rousseau (1990)), the arcs, apparently deviating from (5), are not informetric of nature. Indeed, there are frequently several high-ranking sources that provide the same number of items : there might be a large number of sources with 3 items, a larger number with 2 items and an even larger number of sources with only one item each. Since increases of $R(r)$ at these ranks are linear in r (per group of equal productivity), the graph of R versus $\log r$ is exponential (per group of equal productivity). These exponential graphs get more visible as the groups of

sources with equal productivity get longer. This explains these arcs near the end of a Leimkuhler curve.

Hence this phenomenon is a purely mathematical consequence and has nothing to do with informetric aspects such as (5) or the so-called Groos droop - see Groos (1967).

This first example was a case of mathematical explanation. The graph of Wallace (Fig. 1) was explained with a statistical-probabilistic argument. The next example yields a purely probabilistic explanation of another regularity.

III.2 The relation between the fraction of multinational publications and the fractional score of a country.

In Nederhof and Moed (1993) a slightly different problem as in the title is studied : the relation between the fraction of multinational publications of a country c (i.e. the fraction of publications in which, besides an author of country c also at least one other author appears, belonging to another country $c' \neq c$) and the country's fractionated score. In this scoring scheme a country receives a score $1/b$ in a paper if b is the total number of different countries in this paper and if $b \neq 1$. If $b=1$ the fractionated score is 0. Only in this last element, the fractionated score is different from the fractional score : here the score is 1 if $b=1$.

The regularity found in Nederhof and Moed (1993) is seen in Fig. 3 (reprinted with permission). The concave decrease is clear and is explained in Egghe (1999) but with "fractionated" replaced by "fractional" (leaving open the other problem). Using the fractional score of a country c , it is possible to determine the fraction of multinational publications of c , using elementary techniques of probability theory (such as independence and conditionality). The obtained relation between the fraction of multinational publications of c and its fractional score is a certain average of functions of the type

$$y = 1 - \frac{x^a}{1 - (1-x)^a} \quad (6)$$

where $a \in \mathbb{N}$ and $x \in [0,1]$. All of these functions are concavely decreasing.

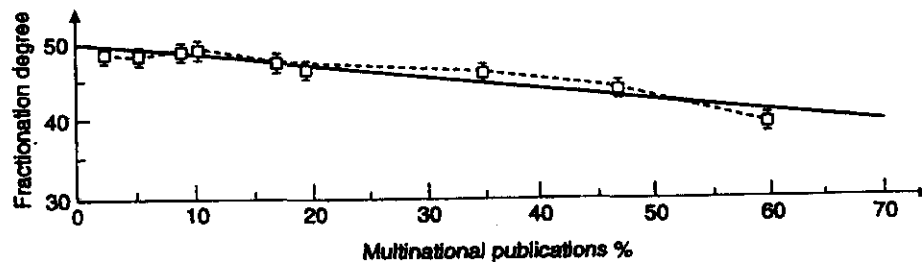


Fig. 3 The relation between the fraction of multinational publications of a country and its fractionation degree.

The next example deals with the relation between the Price Index and the mean or median reference age. This last variable is - essentially - the same as the citation age (i.e. the synchronous one) as studied in the previous section. Also in this case, no clear graph but a cloud of points is found (see below). But, interestingly, here an informetric - probabilistic explanation is needed, contrary to the case in the previous section.

III.3 The Price Index and its relation between the mean and median reference age.

In a fixed literature set, the Price Index PI_d is the proportion of the references that are to the last d years of literature. Price (1970) uses $d=5$, Glänzel and Schoepflin (1995) use $d=2$. In this last paper, the relation between PI_2 and the mean reference age is studied. A graph as in Fig. 4 is obtained (reprinted with permission).

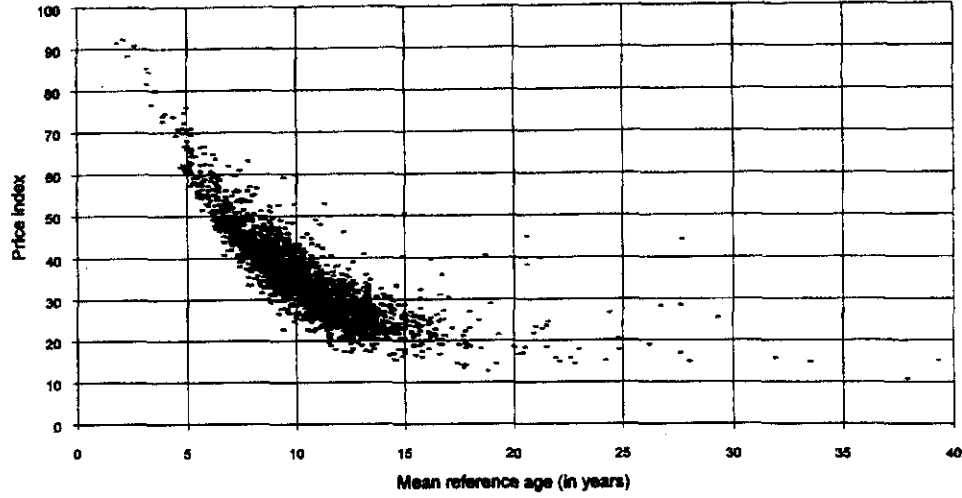


Fig. 4 Plot of the Price Index ($d=2$) versus mean reference age

At first glance, it resembles a bit the graph in Fig. 1 : in both cases we have a cloud of points (rather than a clear graph of a function - as e.g. in III.2) and in both cases the overall impression is that the graph is convexly decreasing. Of course here, not the entire first quadrant, below a certain decreasing function, is filled-up as is the case in Fig. 1.

We can report here that both regularities are completely different in nature and that they have completely different explanations. The explanation given for the Price Index versus the mean (and also for the median) reference age is given in Egghe (1997) and is based on the relation

$$PI_d = \int_0^d c(t) dt , \quad (7)$$

where c is the age distribution of the references. Using for $c(t)$ the exponentially decreasing distribution

$$c(t) = d \cdot a^t , \quad (8)$$

where d and a are constants and $0 < a < 1$, (7) yields an explanation for the convexly decreasing aspects but not for the cloud of points, with the apparent thickness in the middle part. The explanation for this is obtained using the more realistic lognormal distribution

$$c(t) = \frac{1}{t\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2}, \quad (9)$$

where $\ln t$ denotes the Neperian logarithm and μ and σ are constants denoting the mean and standard deviation of $\ln t$. This distribution itself is used and explained e.g. in Egghe and Rao (1992a) and is now generally accepted as age distribution of references (or citations). The fact that it has one more parameter than (8) yields the explanation of the cloud of points - see Egghe (1997).

We hence have here an informetric-probabilistic explanation and hence this regularity has more informetric value than the ones described above in the sense that the latter belongs more to the "informetric theory" than the ones above. Yet, in none of the above arguments, the classical laws of informetrics (Lotka, Bradford,...) are used to explain regularities. This will be the case in the next explanation.

III.4. An explanation of the first-citation distribution.

In a fixed literature set one can look - for each article - at the time t_1 at which this article receives its first citation. Here t_1 is interpreted as "time after publication of the article". Over the whole literature set one can then wonder what is the underlying first-citation distribution. Based on data collected in Gupta and Rousseau (1999), Motylev (1981) and Rousseau (1994) we know already that the cumulative first-citation distribution can be of two types :

- (i) concavely increasing
- (ii) S-shaped : convexly increasing followed by a concave increase.

Rousseau (1994) tries to explain these regularities by determining two differential equations, which yield two cumulative distributions, fitting each type of curve very well. As noted in Egghe (2000) the drawbacks of such a methodology (although it is better than only statistical fitting) are

- two different rationales are needed in order to “explain” both models
- such types of “explanation” start from unexplained dynamics (through the simple formulation of the differential equations)
- they do not involve any previously established informetric results.

In Egghe (2000) we proposed a model in which we use the simple decreasing exponential distribution (8) for the age distribution of the citations in combination with Lotka’s law

$$\varphi(A) = \frac{D}{A^\beta}, \quad (10)$$

where A is the number of citations per article and $\varphi(A)$ is the fraction of articles with A citations. This combination yields for the cumulative first-citation distribution $\Phi(t_1)$ the following formula :

$$\Phi(t_1) = \gamma(1-a^{t_1})^{\beta-1}, \quad (11)$$

where γ is a constant. In Egghe (2000) one can see that (11) is capable of fitting both types of first-citation distributions and that the fits are very good. Moreover the case $1 < \beta \leq 2$ takes care of the concave case and $\beta > 2$ takes care of the S-shaped case. This link between shapes of first-citation distributions and Lotka’s exponent β was new and gives intrinsic informetric explanations. In this sense we consider the discovery of (11) - amongst the other explanations reviewed in this paper - the one with the highest informetric value.

III.5 Concluding remark.

We note the importance of detecting regularities in graphs of (clouds of) points. We also underline the importance of giving a rationale for these regularities and furthermore to

determine if the rationale is informetric in nature or not. If so, the link with previously known informetric distributions must be established which gives then - in turn - a deeper explanation of the observed phenomena.

We want to stimulate the reader to find other regularities in existing literature or (which is of course more difficult) to find new ones. Then we pose the open problem of explaining them (at least partially) in the sense described above.

References

- A. Bookstein (1977). Patterns of scientific productivity and social change : a discussion of Lotka's law and bibliometric symmetry. *Journal of the American Society for Information Science*, 28, 206-210.
- A. Bookstein (1990a). Informetric distributions, Part I : Unified overview. *Journal of the American Society for Information Science*, 41(5), 368-375.
- A. Bookstein (1990b). Informetric distributions, Part II : Resilience to ambiguity. *Journal of the American Society for Information Science*, 41(5), 376-386.
- S.C. Bradford (1934). Sources of information on specific subjects. *Engineering*, 137, 85-86. Reprinted in : *Collection Management*, 1, 95-103 (1976-1977). Also reprinted in : *Journal of Information Science*, 10, 148 (facsimile of the first page) and 176-180 (1985).
- B.C. Brookes (1973). Numerical methods of bibliographic analysis. *Library Trends*, 22, 18-43.
- L. Egghe (1989). The duality of informetric systems with applications to the empirical laws. Ph. D. Thesis. The City University, London (UK).
- L. Egghe (1990). The duality of informetric systems with applications to empirical laws. *Journal of Information Science*, 16(1), 17-27.
- L. Egghe (1997). Price index and its relation to the mean and median reference age. *Journal of the American Society for Information Science*, 48(6), 564-573.
- L. Egghe (1999). An explanation of the relation between the fraction of multinational publications and the fractional score of a country. *Scientometrics*, 45(2), 291-310.
- L. Egghe (2000). A heuristic study of the first-citation distribution. *Scientometrics*, to appear.
- L. Egghe and I.K. Ravichandra 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. Ravichandra 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 (1996a). Average and global impact of a set of journals. *Scientometrics*, 36(1), 97-107.
- L. Egghe and R. Rousseau (1996b). Averaging and globalising quotients of informetric and scientometric data. *Journal of Information Science*, 22(3), 165-170.
- 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*; June 7-10, 1995, River Forest, Il. (USA), 177-185, Learned Information, Medford, N.J. (USA).
- B.M. Gupta and R. Rousseau (1999). Further investigations into the first-citation process : the case of population genetics. *Libres* 9(2), aztec.lib.utk.edu/libres/libre9n2/fc.htm.
- O.V. Groos (1967). Bradford's law and the Keenan-Atherton data. *American Documentation*, 18, 46.
- G. Herdan (1960). *Type-token Mathematics. A Textbook of mathematical Linguistics*. Mouton, 's Gravenhage.
- A.J. Lotka (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323.
- B. Mandelbrot (1954). Structure formelle des textes et communication. *Word*, 10, 1-27.
- B. Mandelbrot (1977). *The fractal Geometry of Nature*. Freeman, New York.
- V.M. Motylev (1981). Study into the stochastic process of change in the literature citation pattern and possible approaches to literature obsolescence estimation. *International Forum on Information and Documentation*, 6, 3-12.
- A.J. Nederhof and H.F. Moed (1993). Modeling multinational publication : Development of an on-line fractionation approach to measure national scientific output. *Scientometrics*, 27(1), 39-52.
- P. Praunlich and M. Kroll (1978). Bradford's distribution : a new formulation. *Journal of the American Society for Information Science*, 29, 51-55.

- Price, D. De Solla (1970). Citation measures of hard science, soft science, technology and nonscience. In : C.E. Nelson, D.K. Pollack (eds.), *Communication among Scientists and Engineers*, 3-22, Heath, Lexington, MA (USA).
- Price, D. De Solla (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.
- R. Rousseau (1994). Double exponential models for first-citation processes. *Scientometrics*, 30(1), 213-227.
- E.G. Summers (1983). Bradford's law and the retrieval of reading research journal literature. *Reading Research Quarterly*, 19, 102-109.
- D.P. Wallace (1986). The relationship between journal productivity and obsolescence. *Journal of the American Society for Information Science*, 37(3), 136-145.
- K.S. Warren and V.A. Newill (1967). *Schistosomiasis, a Bibliography of the World's Literature from 1852-1962*, Western Reserve University, Cleveland, Ohio (USA).
- E.A. Wilkinson (1973). *The Bradford-Zipf Distribution*. OSTI Report #5172, University College, London, UK.
- G.K. Zipf (1949). *Human Behavior and the Principle of least Effort*. Addison-Wesley, Cambridge. Reprinted in 1965, Hafner, New York.