INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

IMPACT OF RESEARCH PERFORMANCE AS MEASURED BY CITATIONS : A NEW MODEL

A.F.J. van RAAN

Science Studies Unit, LISBON-Institute, University of Leiden, Stationsplein 242, 2312 AR Leiden, The Netherlands

Abstract

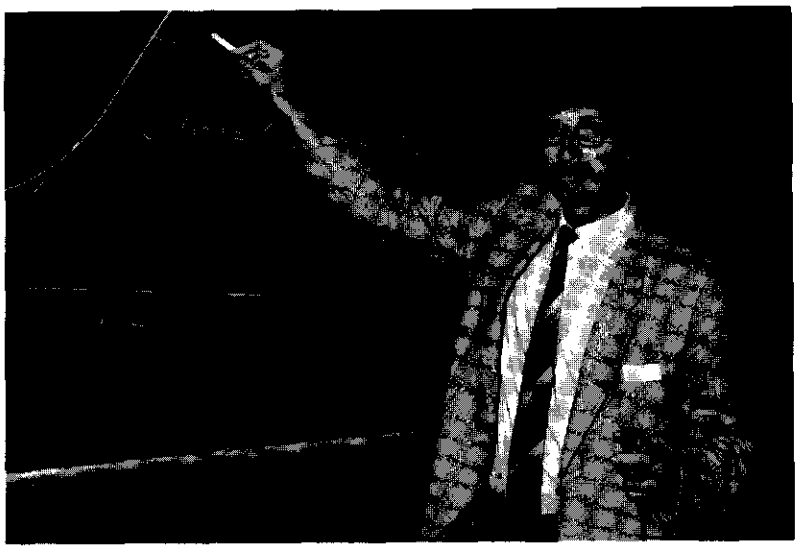
By analyzing the distribution of publications as a function of received citations, a relation was found between the deciles of this distribution and the number of citations. This relation suggests a model analogous to Beer's Law in physics on the absorption of light in a homogenous layer of absorbing material. The results show a remarkable constancy over time. A description of this new model and its consequences on citation behaviour are presented in this paper.

1. INTRODUCTION

In a foregoing paper a method is described to visualize the "income distribution" of journals or any other defined set of scientific publications in terms of received citations in a certain period of time (impact), with differentation to specific types of publications. We introduced the concept of "decilotopes" as a time-dependent representation of the deciles (top-10%, top-20%, ..., lowest-10%) of the income distribution, and to establish a method in order to compare the impact of an *individual publication* with *its* "*environment*" (*e.g., the journal* concerned).
In this study we focus on the relation between the deciles and the number of citations and its behaviour as a function of time. The data strongly suggest a relation with a form analogous to Beer's Law in physics : an exponential decrease of the intensity of light beam passing through an absorbing layer.
From this analogy a model of impact gain (or : loss) is constructed on the basis of the assumption that the number of received citations is a valid representation of the usefulness of the publication concerned for the readers (colleague-scientists). In this model, then, the number of publications in the collection from which the readers choose their references (*i.e., publications cited* by them) play a crucial role.

2. METHODS AND TECHNIQUES

For the discussion of the collected data (all publications in the 1978-volumes of the journal Science, and the citations received by these publications in the years 1978-1983, thereby operating on 1,639 publications and 27,892 citations) and of the method and techniques to handle these data and to construct graphical displays of impact characteristics, we refer to our recent paper (Van Raan and Hartmann, [1]). More specifically, we focus on fig. 2 of that paper, where the so-called decilotopes of the journal "income distribution" are plotted as a function of time. This figure is presented here again as fig. 1. A decilotope can be denoted as $N_C = f(n,T)$, indicating the maximum number of citations per article in the bottom of the percentile n of the distribution concerned (n is percentile 90) in a specific (citing) year T. With this example, we see in fig. 1 that for



A.F.J. Van Raan

Science-1978 publications, in the year 1981, 90 % of all publications (n = 90) had N_C = 15 or fewer citations. We define $\Delta n = 100-n$; for example, for n = 80%, Δn = 20, etcetera.
From the data presented in fig. 1, the graphs in fig. 2 are derived. Here on the ordinate the percentile n is used as a variable, on the abcissa we have N_C. To help the reader, we indicated the 1980-values in fig. 2 with A, B, ..., E (for the decilotopes with n = 90, 80, ... 50, respectively), these points correspond to the points A, B, ..., E in fig. 1.

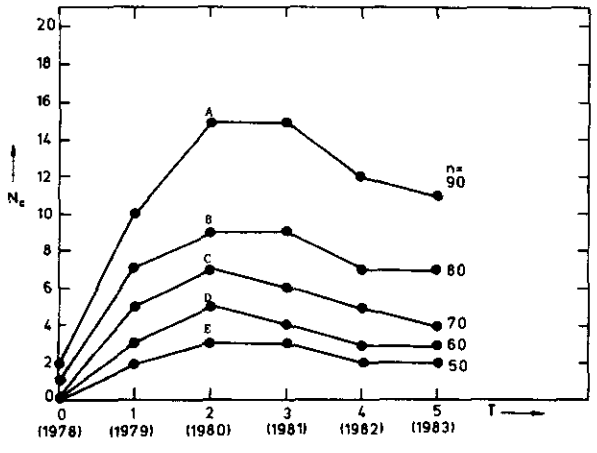


Figure 1 : Decilotope graph for science 1978.
T = number of years after publication (publication year 1978 is T=0).
N_C = maximum number of citations per article for the bottom n % of the article distribution (see text).
n = the numerical value of the decilotope.
A,B,C,D,E are markers to indicate corresponding points in figure 2.

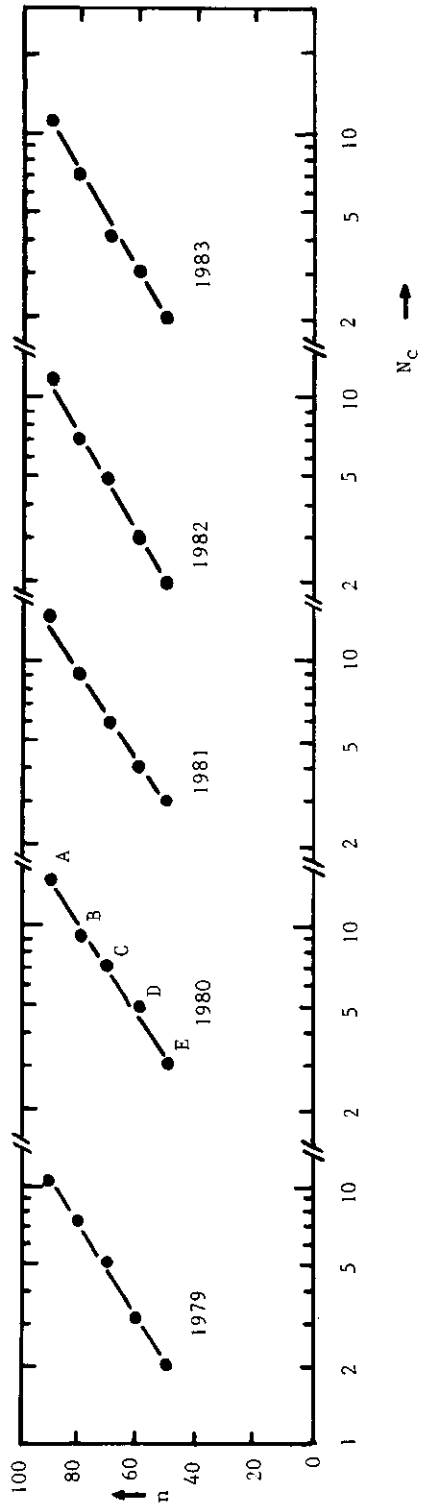


Figure 2 : Decilotope-Citation Distribution for Science 1978 calculated for five successive years (1979-1983).
N_C = number of citations as specified in fig. 1 (repeating log-scale).
n = the numerical value of the decilotope.
A,B,C,D,E are markers as explained in Figure 1.

3. MODEL AND DISCUSSION OF THE RESULTS

The results of fig. 2 are quite surprising. First, from a very skewed and in principle unknown "income distribution" of a journal, we now find a log-lin relation between n (the percentile of the distribution-function) and N_C (the maximum number of received citations), and therefore between Δn and N_C :

$$n(N_C) = n(1) + \gamma \log N_C \qquad (1)$$

giving

$$\frac{dn}{dN_C} = \gamma N_C^{-1}, \qquad (2a)$$

which is in fact the number of publications $P(N_C)$ having exactly N_C citations, so that

$$P(N_C) = \gamma N_C^{-1}. \qquad (2b)$$

Eq. (2a) can be written as

$$\frac{dN_C}{N_C} = \gamma^{-1} dn, \qquad (3)$$

yieldings after integration between an arbitrary value of n and n = 100 :

$$N_C = N_C^{\circ} \exp(n/\gamma) = N_C^{\circ} \exp((100 - \Delta n)/\gamma) \qquad (4a)$$

where N_C° is the extrapolated value of N_C for n = 100, i.e. $\Delta n = 0$.

We now rewrite eq. (4a) as

$$N_C = C.\exp(-\beta.\Delta n), \qquad (4b)$$

with $C = N_C^{\circ} \exp(100\beta)$, and $\beta = \gamma^{-1}$.

A second interesting finding illustrated by Figure 2 is the virtual constancy of the gradient γ (0.60 ± 0.03) for the successive citing years T (1979-1983). This means, according to eq. (2b), that the number of publications having N_C citations, $P(N_C)$, has a constant proportionality over time with the inverse number of citations (N_C^{-1}). We hypothesize that this constant proportionally (the gradient γ) is a characteristic of the set of entities under study, i.e. a set of scientific articles in the form of a journal, more specifically, the volumes of one particular publication year for that journal (Science, 1978).
Let us discuss the empirical findings. First, the log-lin relation between n and N_C.
We notice the formal similarity of eq. (1) with a simple mathematical equation of Bradford's [2] law as given by, for example, Narin [3]. In a very recent paper, Chen and Leimkuhler [4] derive a common functional relationship among Bradford's law and the two other well-known empirical laws of information science : Lotka's law [5] (originally on scientific productivity) and Zipf's law [6] (originally on word frequency). We are further investigating the relation between the citation-distribution function discussed here and these well-known laws. In this paper however we focus on the construction of a model to explain our empirical findings.
In eq. (4b) C and β are constants to be derived from the empirical results. This equation in fact models the empirical income distribution. On the basis of the formal structure, we suggest an analogy of Beer's law in physics (classical optics) :

$$I = I_0 \cdot \exp(-\alpha \Delta l), \qquad (5)$$

where I is the light intensity of a beam with initial intensity I_0 after passing through an absorbing layer with thickness Δl and with absorption coefficient α. The equation results from the assumption that the absorbed light (ΔI) is -in first approximation- proportional to the incoming radiation intensity and the thickness of the absorbing layer

$$\Delta I = - I.\alpha.\Delta l \qquad (6)$$

which yields after integration eq. (5) as a solution.

The analogy with Beer's law means that the impact of a publication can be compared with radiation shining through an absorbing layer : the thickness of this absorbing layer is Δn, i.e. the top-layer of the distribution. In other words, the eventually received impact is a result of an absorption process wherein the "initial radiation" (in our analogy : the hypothetical initial or ideal, "undisturbed" impact, given by parameter C in eq. 4b) is absorbed by a part of the total amount of publications. For example, when n = 60, the "absorbing layer" is the top-40% of the distribution ($\Delta n = 40$) : a thick layer, which results in a rather small outcoming radiation or, in our case, rather small received impact. Thus, what we call "gaining impact" would in fact be, in this model, a loss of the hypothetical initial impact induced by competing other publications ("the absorbing layer"). For the case we have a publication in the top-10% ($n = 90, \Delta n = 10$), the "absorbing layer" is very small, and the "loss of initial impact" will be small too, resulting in a high "received" impact. The consequences of this analogous modelling are threefold.

First, it must be assumed that colleague-scientists (readers) discriminate on qualitative grounds between the values of publications. One could argue, that a large panel of peers judged the usefulness for their (the reader's) own purpose of -in our example- the Science-1978 publications and in this way a ranking or grouping of these publications emerged, giving rise to top-10%, top-20%, ..., lowest-10% publications, in terms of usefulness for the Science-1978 readers. Because the number of Science readers ("the audience" for Science) is very large, this assumption is on statistical grounds quite plausible. In fact, there is a strong resemblance with the voting behaviour of the public in elections.

Second, it must be assumed that this "intersubjective" or "collective" judging of usefulness is represented in a statistically significant way by the numbers of received citations. Then, the "intersubjective" top-10%, top-20%, ..., lowest-10% correspond with the top-10%, top-20%, ..., lowest-10% as measured by citations.

Third, as a consequence of the foregoing points, it must be assumed that Science-publications "compete with each other" for usefulness : it is the "thickness" of the absorbing layer (Δn), i.e. a specific top-decile(s) collection of Science-publications, which determines the "loss of usefulness".

Altogether, we hypothesize that colleague-scientists (i.e., readers of scientific publications in their field of research) compose their reference-lists by using "sources", mostly scientific journals, but also conferences, private communications, etcetera. From these sources, publications can be used as a reference. This use as a reference then is, in our model, primarily determined by competition of publications within the source, in our example within Science-1978. So the use of a specific Science-1978 publication, reflected by a reference to this publication, primarily depends on its "usefulness" compared with the other -competing- publications in Science-1978. Of course, for the readers (colleague-scientists) involved, Science must be a main source, and not, which is for example the role of Science for a field like physics, an additional source or a source of "general information". The above model is rather drastic : it means that the process of receiving citations is mainly determined by competition of publications within a main source. Decisive for this competition is the intersubjective ranking (top-10%, top-20%, ..., lowest-10%) of the publications in this source according to their usefulness for the reading colleague-scientists. In general, a "main source" can also be a group of field-specific journals.

The second empirical finding was the constancy of the gradient as a function of time. The consequences of this finding are that with a given initial set of scientific articles, the position of the articles in citation-classes as defined by top-10%, top-20%, etc., remains more or less the same over a period of, at least, 5 years. Or, in other words : when an article is within a specific citation-class in the first year after publication, then there is a high probability that it remains in the same class (or : will keep its relative impact position) for the next four years. Of course, our empirical findings only regard citation-based impact measures of articles published in the journal Science, and in the year 1978. On the other hand, it is hardly plausible that our findings would only be the result of some very peculiar characteristics of just Science 1978. Therefore, our conjecture is that we found a more general empirical fact. Work is going on to investigate the reported findings for other journals and various publication years.

5. CONCLUSIONS AND COMMENTS

On the basis of empirical findings with a dataset of 27,892 citations in the period of 1978-1983 to 1,639 publications appeared in the 1978 volumes of the journal Science, we constructed a model of gaining impact with the following elements :

1. Scientists establish collectively a ranking of publications within a journal like Science according to the usefulness of these publications. We may compare this process with elections of political parties (for example in the European political system) : a top-10%, top-20%, ..., lowest-10% grouping of publications is achieved.
2. This judgement of usefulness is represented in a statistically significant way by numbers of received citations. Thus, the "intersubjective" top-10%, top-20%, ..., lowest-10% correspond in reasonable approximation to the top-10%, top-20%, ..., lowest-10% as measured by citations.
3. Reference lists of scientists are assumed to be composed of references from specific sources. As soon as a source can be regarded as a main source of information for a field of research, scientists "reserve" in fact a part of their references list for publications from that source(s).
4. Publications within such a main source (e.g., a large journal, or a set of field-specific journals), have to compete with each other for their usefulness. The results of this competion are dictated by the intersubjective ranking as formulated under point 1.
5. Statistically, this top-10%, top-20%, ..., lowest-10% layer model works as follows : the initial usefulness (the maximum would be that every reading colleague-scientist gives one citation to a specific publication) decreases by "absorption" of (parts of) this initial usefulness as a result of the competition of other publications in the layer (e.g. top 20%) concerned. To describe this model, we use the analogon of light (with a fixed, initial intensity) passing through an absorbing layer.
6. At least for a highly topicality-oriented journal like Science, the position of an article in terms of received impact (citation class) established in the first year after publication, will have a high probability to remain the same over a period of at least five years after publication.

REFERENCES

[1] van Raan, A.F.J. and Hartmann, D., The Comparative Impact of Scientific Publications and Journals : Methods and Graphical Display. Scientometrics (1987) 11, pp. 321-327.

[2] Bradford, S.C., Sources of information on specific subjects. Engineering (1934) 137, pp. 85-86.

[3] Narin, F., Evaluative Bibliometrics : The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity, National Science Foundation, Washington (1976) Monograph pp. 456, NTIS Accession nr. PB 252339/AS.

[4] Chen, Y.S. and Leimkuhler, F.F., A Relationship Between Lotka's Law, Bradford's Law, and Zipf's Law, Journal of the American Society for Information Science (1986) 37, pp. 307-314.

[5] Lotka, A.J., The frequency of distribution of scientific productivity, Journal of the Washington Academy of Science (1926) 16, pp. 317-323.

[6] Zipf, G.K., Human Behaviour and the Principle of Least Effort, (Cambridge, MA : Addison-Wesley, 1949).