INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)

GROWTH AND CITATION SELECTION RATES IN RAPIDLY GROWING SCIENCES FROM DATE STACKING AND BIBLIOGRAPHIC DATABASES

D.H. HALL

Department of Geological Sciences, The University of Manitoba
Winnipeg, Manitoba R3T 2N2 (Canada)

Abstract

Growth of literature in a field was taken as an indicator of the growth of the field as a whole. Stacked citation data and the yearly accumulation of literature were found for 3 fields in the geosciences : Geology and Geophysics of Mars (1962-1985), Magnetic Stratigraphy (1957-1977) and early Seismology (1600-1750). The yearly accumulation for the first two fields was found from a computerized bibliographic database (GEOREF). That for early Seismology was found from a published bibliography. Phases of logistic growth and of exponential growth as well as several periods of constant rate of selection of citations were found.

INTRODUCTION

CITATION ANALYSIS

The scientific paper, developed during the Scientific Revolution, has continued almost unchanged to the present day. Principal features of this form of scientific communication are that a scientific paper is usually confined to an easily definable subject field and refers back to previous literature in that field. The paper itself becomes the point of departure for further papers. The lists of References (or Citations) which are appended to scientific papers reveal much about the nature of science and scientists. There has arisen a whole field within the science of science built about an analysis of the way in which authors select their references. It has been known for the past 60 years (see for example De Solla Price [1]) that the rate at which the number of citations in a paper drops off with their age relative to the date of the paper is a measure of the rate of growth of the subject field. Citation analysis has expanded since the time of its discovery and is applied to many fields such as in mapping networks of ideas using co-citation analysis, a widespread application of citation analysis at the present time.

AGES OF CITATION

The analysis of ages or dates of citations has a unique and important place in the study of science for a number of reasons. One is that such analysis is an indicator which is frequently used in science policy studies [2]. Studies of this type could benefit from further research on citation age or date analysis. This type of analysis is the only way to measure the rate of growth of a science during periods for which computerized bibliographic databases (or good hardcopy databases) do not exist. Thus the method is of potential importance in studies of the history of science or in any situation where bibliographic data are poorly known. In addition, this type of analysis has the potentiality of

focussing on short periods of time, by using source papers whose citation data intersect the required period.

The reason that authors of scientific papers, unwittingly, reveal facts regarding the growth of their subjects is explained as follows.

GENERAL STATEMENT

In selecting references (which we will henceforth call *citations*) to use and list in a paper, an author chooses from the total accumulation of bibliometric materials in or close to the field in which the paper falls. If the citations are used for citation indexing or for citation analysis, the paper is referred to as a *source paper*. Let the date of the source paper be t = S, and that of a citation be t = S - T. Thus T represents the age of the citation with respect to the source paper. The date D of the citation is then given by t = D = S - T, and its age by T = S - D. Authors tend to select citations according to age : the younger the citation the more likely it is that it will be selected [4]. The selection is made at a rate r(T) = r(S-D) from N'(D) items, the pool of literature in the field for the year t = D, corresponding to the age T. Thus the number of citations of age T is given by

$$C(T) = C(S-D) = r(S-D)\, N'(D) \qquad (1)$$

The quantity r(S-D) is the author's rate of selection, and represents the probability that a paper published in the year t = D will be chosen as a citation by an author at the future time t = S. N'(S-T) represents the number of bibliographic items for that year. We know that [3]

$$N'(S-T) = \left.\frac{dN}{dt}\right|_{t=S-T}, \qquad (2)$$

where N is the total accumulation up to the year t = S - T.

RATE OF CHOOSING CITATIONS

De Solla Price [4] describes a remarkable instance in the growth of science in which the literature grew exponentially while citation numbers decayed exponentially at an identical rate implying that the selection rate remained constant over the time period under study. He noted that "this rather surprising result" means that for the lifetime of the growth phase studied a publication would have a constant chance of being used as a reference in a future paper no matter what its age [4]. We have no reason to expect to find such a remarkable behaviour of r(T) always, but the example suggests that many instances of interesting trends in selection rate can be found. The understanding of the magnitudes and changes in r(T) for various growth models is important because eq.(1) tells us that given knowledge of r(T), the growth of the literature in a field can be found from citation numbers in a collection of source papers. This could have application in historical studies of science, for which comprehensive bibliographies cannot be compiled.

Rapidly growing fields, found in all areas of science, are perhaps the simplest examples of growth. In the present paper, examples from the geosciences will be used. Menard [5] found that rapidly growing areas are good cases for study in the geosciences. For these reasons two collections of source papers from rapidly growing fields in geoscience : Geology and Geophysics of Mars, and Magnetic Stratigraphy, were chosen for study in this paper. The former is important in geoscience because of the linking of many fields within it to the upsurge of planetary science in the past quarter century. The latter, Magnetic Stratigraphy, is a principal element in a recent rapid advance in geochronology

which arose from the merging of techniques and ideas from rock and terrestrial magnetism with geological and radiometric geochronologic methods and ideas. In addition, an early period in seismology was examined. The results from the fields studied in the present paper lead us to consider two types of growth, logistic and exponential.

GROWTH MODELS FOR SCIENTIFIC FIELDS

Growth of science and scientific literature.
To characterize the growth of a science fully a number of different indicators must be used. These reflect such things as economic factors, scientific ideas, scientific manpower, infrastructure, and the output of scientific literature.

It has been found that these indicators are often correlated with each other and with the overall growth of the science. The output of scientific literature has proven to be a good representation of the growth of science as a whole [1] and of growth of the geosciences, the subject of the present paper [5,6]. The growth of literature in three fields is the subject of the present paper, and this growth will be taken as representative of the overall growth of these fields. This assumption whould, however, be tested as often as possible. Many different types of growth of sciences have been identified and quantified in past scientometric research. In the present case two types, logistic and exponential, are found. The basic equations describing these modes of growth are outlined in the following sections.

Estimation of growth by non-citation methods.
Growth can best be measured from computerized bibliometric databases, and these give estimations of time series of dN/dt. The techniques used in the present paper are fully described in earlier papers [3,7]. The accumulation of literature, N(t) can be found by numerical integration of dN/dt. In most cases the simplest form of integration, cumulation, is sufficient. If computerized databases do not cover the time period of interest, then hardcopy bibliographies and indices (not usually a source that can be used to full advantage) must be employed, or citation analysis. Let us now examine the features of exponential and logistic growth.

EXPONENTIAL GROWTH

We write, for an exponential growth phase in a field identified as beginning at a time t = 0,

$$N(t) = N_0 + ae^{bt} \qquad 0 \leq t \leq t_m, \text{ and } b > 0 ; \tag{3}$$

N_0 is the accumulation of the field at t = 0, the beginning of the exponential phase, and t_m is the time span treated.

From equation (3) we have :

$$dN/dt = abe^{bt} \text{ , where } b > 0 . \tag{4}$$

Using equation (4), equation (3) can alternatively be written as :

$$N(t) = N_0 + 1/b \; dN/dt . \tag{5}$$

DOUBLING TIME

In equations containing exponential terms $e^{\pm bt}$ the time, t_d, required for N or

N' to double or to decay to half depending on the sign of b is given by :

$$t_d = \frac{\ln 2}{|b|} \tag{6}$$

This is a special case of the equation for t_k, the time for N or N' to grow or decay with a factor k :

$$t_k = \frac{\ln k}{|b|} \tag{7}$$

LOGISTIC GROWTH

For logistic growth [9,6] if y represents the cumulative growth and dy/dt = y', approximately its growth over a short interval (such as one year) we write :

$$y = \frac{K}{1 + \beta e^{-\alpha t}} \tag{8}$$

and y' by :

$$y' = \alpha y(1-y/K) \tag{9}$$

y' is a symmetrical curve peaking at :

$$t = 1/\alpha \ln \beta \tag{10}$$

The height of the peak is given by :

$$y'_c = \alpha K/4 \tag{11}$$

The peak in y' occurs at the inflection (or critical) point in y (y_c), the point of most rapid growth of the logistic. At this point, y is given by

$$y_c = K/2 \tag{12}$$

The doubling time for a logistic growth phase can be estimated as follows :

$$t_d = y'_c/y_c = 2/\alpha \tag{13}$$

Equation (13) has an important consequence, to be seen in a later section.

Approximation to exponential growth.
There are two intervals in a logistic growth phase which approximate exponential growth. These are given by

$$t < t_c - 1/\alpha \quad \text{and} \quad t > t_c + 1/\alpha \quad \text{for y .} \tag{14}$$

In the first,

$$y \approx K/\beta e^{\alpha t} \tag{15}$$

and in the second

$$y \approx K(1 - \beta e^{-\alpha t}) \ . \tag{16}$$

A 1/e criterion is used to define the limits in eq.(14). Thus a logistic starts to grow approximately exponentially, and approaches the logistic limit y = K approximately exponentially in the latter part of the growth, with a central "window" bearing no resemblance to the exponential mode of growth.

Conformity of growth of a field to a model.
Even if a field as a whole conforms to some particular model of growth, citations from individual source papers cannot be expected to adhere exactly to it. Authors choose the references that they need to support their papers and to acknowledge the sources of ideas and results they have used. Eq.(1) ensures that authors will choose references with ages that tend to bear some relationship to the prevailing growth model. But the specific needs of each paper will mean that for some ages citation numbers will be out of proportion to the numbers expected from equation (1). Of course these needs will vary from author to author. Thus we can consider the actual selection of citations by authors to be perturbed from the model by random deviations giving the sum of a "model" and a "random" component. If $M(S,T)$ is the number of citations for age T expected from the model for a given source paper dated $t = S$, and $\delta(S,T)$ is the random deviation from the model for that particular source paper,

$$C(S,T) = M(S,T) + \delta(S,T) \qquad (17)$$

STACKING

From the above discussion we can see that a collection of source papers is required to establish the growth model for a particular field. This requirement arises because some sort of statistical combination of source papers is required to separate the (constant) model component from the random deviation component in equation (17). These two components represent signal and noise respectively. Stacking is a procedure which is commonly used in signal processing in fields such as physics, geophysics, and engineering to enhance the signal to noise ratio in multichannel data, and it proves to be a simple and effective procedure to extract information on growth of a field from the citations in a collection of source papers.

The stacking process can be illustrated by the array shown in Table 1. Suppose we have a collection of source papers numbered $1,2,\ldots,i,\ldots,L$, with dates given by $S_1,S_2,\ldots,S_i,\ldots,S_L$. Let each of these papers have citations of age $T = 0,1,2,\ldots,$yrs. The number of citations of age T_j for the i^{th} source paper is given by $C(S_i - T_j)$.

Each row in Table 1 represents the citations in a source paper. Each row contains the signal (M(T)), which represents the model, and noise $\delta(S,T)$ as given in eq.(17). Each of these rows is a "trace" or a "channel" in the usual terminology of signal processing. Thus the list of citations in a paper can be thought of as a noisy channel through which the growth model for the field in which the paper lies is transmitted. The citation method for estimating growth of a science can thus benefit from techniques used in signal processing such as stacking and calculation of channel capacity. Stacking of the traces enhances the signal in comparison to the noise. Stacking may be done in two ways, and these are detailed as follows.

In the best case, a straight-sum (SS) vertical stack can be used [8]. This case occurs if we have coherent signal (M(S,T)) and gaussian noise ($\delta(S,T)$). In this case of ideal statistics we know that if source papers are stacked, a signal/noise improvement of $\sqrt{K}$ can be achieved. Before stacking is carried out, the statistics of each trace should be evaluated so that the degree of deviation from optimum conditions for stacking can be evaluated. Techniques are well developed in stacking to improve performance when non-optimum data are used [8].

Date stacking.
In this type of stacking we choose a time line (with constant date D) through the array (Tables 1 and 3) joining citation numbers with the same date. Segments of time lines are shown in Table 1 taking as an example S_1 and S_2 to

be one year apart and in Table 3. Suppose there are K citation numbers on this path. A data stack is the normalized sum (arithmetic mean) of these citation numbers and is given by :

$$S_1(D) = \langle C(S_i - D)\rangle = \langle r(S_i - D)\ N'(D)\rangle \ , \qquad (18)$$

applying eq.(1) and averaging over all items $C(S_1 - D),\ldots,C(S_i - D),\ldots,C(S_K - D)$ where K varies with D (Tables 1 and 3). The above equation can be written

$$S_1(D) = \bar{r}\ N'(D) \qquad (19)$$

writing $\langle r(S_i - D)\rangle = \bar{r}$ and noting that N'(D) is constant for all values of i. Eq.(19) is valid for any growth model, requiring only the substitution of the expression for N'(D) for that model. For exponential growth, N'(D) is given by eq.(4) and for logistic growth by eq.(9). N'(D) can also be determined from actual data such as from a computerized bibliographic database.

For example, in the exponential case, substitution from equation (4) into equation (19) gives that

$$S_1(D) = ab\bar{r}\ e^{bD} \ (b > 0) \qquad (20)$$

Thus the datestacks grow exponentially with time at the same rate as the field, just as individual papers do. Individual deviations from the model, however, are suppressed in the stack.

Age stacking.
In its simplest form, this operation consists of adding the columns in Table 1. The sum is usually normalized dividing by L, the number of rows or "traces" in the stack. This is called a straight-sum (SS) vertical stack. If normalized it is simply the arithmetic mean of the citation numbers of equal age for a collection of source papers and is given by :

$$S_2(T_i) = \langle C(T_i)\rangle = \langle r(T_i)\ N'(D_i)\rangle \qquad (21)$$

applying equation (1) and averaging over all items $C(T_1),\ldots,C(T_i),\ldots,C(T_L)$, where L is the same for all stacks. Averaging is not along constant data lines (but rather is along constant age lines) as it is for datestacking, and $r(T_i)$ and $N'(D_i)$ cannot be separated as was done to obtain equation (15). Thus, datestacking should, usually, be preferred over age stacking. In the case $r(T_i)$ = const. = r, we have :

$$S_2(T) = r\ \langle N'(D_i)\rangle \qquad (22)$$

from which r can be found.

There may be source collections in which datestacking is not feasible (if the source papers are scattered over a long period of time such as might occur in a historical collection). In that case, it might be possible to invert eq.(21) for the $r(T_i)$; then $r(T_i)$ can be assigned a mean date.

STACK OR ARITHMETIC MEAN?

The stacks were defined as the arithmetic means of items selected from the array in Table 1. Why not just refer to them as means rather than bring in the ideas and terminology of stacking? The reason is that the literature and practice of stacking as a discipline within signal processing is directed towards enhancement of the signal in relation to noise (the signal being the

model M(S-T) in this paper). Techniques such as weighted stacks and "killing" of statistically unsuitable traces before stacking can be imported directly if stacking is used as the method for improving definition of the signal (M(S-T)) carried in the citation data.

RESULTS

MARS COLLECTION

A collection of 229 source papers on the geology and geophysics of Mars was assembled (Table 2). Some of the citation data are shown in Table 3, along with the paths through the data required to perform date stacking and age stacking.

One of the results of the present study is that date stacking is superior to age stacking. Date stacking is shown by equation (19) to be closely related to the rate of accumulation (dN/dt) of literature in a field. The rate of accumulation of literature is shown in Fig.1 and Table 4; the date stack for the Mars collection of source papers is shown in Fig.2 and Table 4.

The data on the growth of the field were obtained from the GEOREF bibliographic database using the CAN/OLE system. The methods of doing this are fully documented in earlier papers [3,7]. The yearly total, dN/dt is shown in Fig.1 from 1962 to 1985. The datestacking was done by the methods described in earlier sections. The time window (1962-1976) was determined by the number and dates of source papers in the collection through the number of items (K) in the stacks, and stacks were rejected when $K < 100$ (for $D < 1962$ and $D > 1976$). The number of items K was over 200 for the period 1969-1975. In addition, N, the cumulation of dN/dt from 1962-1985, is shown in Fig.3 and Table 4.

FIT OF GROWTH MODELS TO THE MARS DATA

The GEOREF data ($dN/dt = N'$) in Fig.1 and Table 4 (shown in solid line) have the appearance of two successive logistic growth phases in escalation. Thus we first investigate a logistic fit to these data. For this purpose, the cumulative of the GEOREF and Datestack data were added in Table 4 and 5 and these are shown in Figs.3, 4, 7 and 8.

Fitting a logistic.
In the data for this paper, either the GEOREF data (N') or the Datestack data ($S_1(D)$) are represented for the fits by y' of eq.(9). The cumulatives of these quantities are represented by y of eq.(8).

A convenient sequence in the fitting process when there are two logistic phases is as follows.

1. Plot y'/y against y. See [9] for the details of this step. From eq.(9) it can be seen that from this plot the parameters α and K, two out of the three which define the logistic (eq.(8)), can be found. If the first phase is a logistic it will appear in the plot as a straight line with negative slope.
2. An exponential phase will appear as a line of zero slope in this plot (see eqs.(3) and (4)).
3. If step (1) indicated a first-phase logistic, the parameter β for that phase can be then found from eq.(10) by identifying the peak of y' for this phase on Fig.1 or Table 4.
4. Then y and y' for this phase of the logistic fit can be calculated from eqs. (8) and (9).

5. Overlap of the logistics was allowed for. The second phase data, corrected for overlap of the first phase forms a new y'/y plot, leading to the calculation of y and y' for the second phase.
6. The calculation for the first phase may require modification for overlap of the second phase.
7. Data modified to account for overlap is referred to as "reduced data". Reduced data portray logistics in escalation as usually defined [1].

Overlap of the logistics.
The best-fit logistics are shown in Figs.1, 2, 5 and 6. Shown are y as given by two overlapping logistic phases, where

$$y' = y'_1 + y'_2 \tag{23}$$

The subscripts refer to the first or the second phase. In the absence of further information, overlapping phases should be taken as the most logical possibility since the effects of forerunners of coming phases and the aftermaths of previous phases are frequently present in science at any period in its history. Only historical studies can decide in a particular case whether or not overlap is the correct model.

EXPONENTIAL GROWTH

A large number of trial best-fits of an exponential to the data (Tables 4 and 5) and to sections of it were made. The algorithm used was based on the assumption of Gaussian noise [9]. Exponentials prove to fit any single section of the data closely, but there was usually a lack of consistency. For exponential growth, all of N, N', $S_1(D)$ and $S_1(D)$-cumulative should have the same doubling time (see eqs.(3), (4), (6) and (20)). This consistency was in general not found, but was found in one case, to be described in the next section.

Mars data.
The Mars GEOREF (N'(D)) form a time series from 1962 to 1985, and these data are best fit by two logistic phases (Table 6).

However, the Datestack window cuts off the Datestack time series earlier than for the GEOREF series, at 1976. This time series defines a first logistic phase, but the remaining values constitute too short a series to determine whether a second phase logistic exists. The y'/y plots for both the GEOREF and Datestack data indicate that an exponential fits the unreduced data for several years after the critical point of the first logistic phase. Thus, an exponential (best-fit to the unreduced data) was used for the Datestack data ($S_1(D)$) for the period 1970-1976. The following result was found :

$b = 0.21\ yr^{-1}$, $t_d = 3.3$ yr, ρ, the coefficient of determination giving goodness of fit on a scale 0 to 1 [10] was 0.64. (24)

This procedure gives us a means of using the Datestack data to estimate the growth rate of the field from 1970 to 1976, the end of the Datestack window.

As a test of consistency, we compare the best exponential fits of (unreduced) y and y' from GEOREF and y from the Datestack.

Fitting an exponential to (unreduced) y from Datestack for 1970-1976, we find :

$$b = 0.19\ yr^{-1},\ t_d = 3.6\ yr,\ \rho = 0.99 \tag{25}$$

For GEOREF data during a similar period (1974-1980) we have for the best exponential fit to y' (unreduced) :

$$b = 0.14\ yr^{-1},\ t_d = 5.0\ yr,\ \rho = 0.66 \qquad (26)$$

and for the best exponential fit to y (unreduced) :

$$b = 0.16\ yr^{-1},\ t_d = 4.3\ yr,\ \rho = 0.99 \qquad (27)$$

Thus there is agreement among results (24) to (27) that the field (geology and geophysics of Mars) after the first logistic phase started to grow at the rate of about :

$$t_d = (3.3 + 3.6 + 5.0 + 4.3)/4 \approx 4\ yr\ . \qquad (28)$$

From the two longer datasets that we have (the GEOREF y and y' data) it appears that the second phase is, in the longer term, a logistic. The exponential fit, however, shows itself to be a useful indicator of the initial rate of growth after the first phase, because there was a fourfold agreement among the doubling times obtained for y and y' for both the GEOREF and Datestack data. It makes a useful estimate of the growth rate given the short span of the Datestack window.

Magnetic Stratigraphy Collection.
A published collection of 35 source papers from the period 1966-1977 on magnetic stratigraphy [12] was used. The rate of accumulation of literature in the field was derived from the computerized bibliographic database GEOREF by the same search procedure as was used with the Mars collection. The rate of growth of the field (dN/dt) is shown from 1959-1977 in Fig.5, and the cumulative growth (N) in Fig.7. The rate of growth as found from datestacking ($S_1(D)$) from 1957-1975 is shown in Fig.6, and the cumulative of $S_1(D)$ in Fig.8. All of these quantities are tabulated in Table 5. The time window for datestacking was determined by the number of items K in the stacks for each year. Stacks were rejected when K was less than 18 (for $D < 1957$ and $D > 1975$). The number of items in a stack (K) was 35 at its maximum.

FITTING GROWTH MODELS TO THE MAGNETIC STRATIGRAPHY DATA

The GEOREF and the Datestack time series (Fig.5, 6 and Table 5) have the appearance of two successive logistic growth phases. For both the data from the computerized bibliographic database (GEOREF) and the Datestack, logistic curves for two successive phases of logistic growthwere fitted, allowing for overlap. The same procedure as for the Mars data was used. These results are summarized in Table 6, and will be discussed and compared with the Mars results in a later section.

As stated in the section on the Mars data, no consistent pattern of exponential fits to the data in Tables 4 and 5 could be found except for results (25)-(28).

AGE STACKS

No consistent exponential fits could be found to the age stacks that were carried out on the Magnetic Stratigraphy data.

RATE OF SELECTION OF REFERENCES

The basic equation for calculating this quantity ($\bar{r}$) is given in eq.(19). It is found by dividing $S_1(D)$, the Datestack, by N'(D), obtained from GEOREF. Errors can arise in N'(D) and in $S_1(D)$ because of factors such as minor delays before and in publication. These can cause errors in these quantities. It was found in earlier researches [3,7] that a three-year window provides an effective smoothing of this effect. Thus we should modify eq.(19) as follows :

$$\hat{S}_1(D) = r\,\hat{N}'(D) \qquad (30)$$

where the $\wedge$ represent the application of a (1/4, 1/2, 1/4) smoothing operator. The calculated values of r for the Mars collection and the Magnetic Stratigraphy collection are given in Table 7. The significance of these results will be discussed in a later section.

NOTE ON PLOTTING DATESTACK RESULTS

From equation (19) we note that the rate of growth of the field, N'(D) is given by $1/\bar{r}\ S_1(D)$. When $\bar{r}$ is constant the Datestack results will portray N'(D), the growth of the field. Thus the Datestack should be plotted against D, the date for the time line for the stack through the citation data.

COMPARISON AND DISCUSSION ON RESULTS, MARS AND MAGNETIC STRATIGRAPHY DATA

GROWTH PHASES

Mars data.
These results are tabulated in Table 6. Both the data from the computerized bibliographic database GEOREF (dN/dt and N for the field), which gives the total count of papers in the field, and from the datestack of citations ($S_1(D)$ and its cumulative), which gives a selection made by authors from the total count, show that there are two growth phases in succession during the period studied. Logistic curves fitted well (Figs.1, 2 and Table 7) except for the second phase of the datestack, explained in an earlier section. Both dN/dt and $S_1(D)$ indicate that the first phase began about 1962, but they differ on the date of the end of the first phase: the total count (GEOREF) indicates 1975 while the Datestack indicates 1970. As stated in Table 7, the difference could be reduced to three years by a reasonable alternative fit. It was shown in an earlier paper [3] that a three year span about the date of a paper is a fair representation of the uncertainty in the data. Thus the two point to the same span of the phases within allowable limits.

The fact that the data from the bibliographic database (dN/dt) show clearly that there are two logistic growth phases was described in an earlier section. It was also shown that because of its limited time window, the datestack data from citation analysis ($S_1(D)$) define the same first logistic phase defined by the dN/dt data, but does not have sufficient time span to define what sort of growth occurs after that. An exponential growth model was fitted to this second phase for a time span extending into the second phase, exhibiting consistency with three related growth phases.

Magnetic Stratigraphy data.
These results are tabulated in Table 6. The computerized bibliographic database (dN/dt) and datestacking of citations ($S_1(D)$) and their cumulatives have the

appearance of two successive growth phases, possibly logistic curves. The logistic equations fitted well (Figs.5, 6 and Table 6), with provision for overlap.

Both dN/dt and $S_1(D)$ show the first growth phase to have been in progress from the start of their time windows. The end of the first phase was indicated as 1970 for dN/dt and 1965 for $S_1(D)$. As in the Mars data a reasonable alternative fit would reduce the difference to 3 years, an allowable limit given the estimated uncertainty in this type of bibliographic data. A second logistic growth phase follows for both dN/dt and $S_1(D)$, as seen in Figs.5 and 6 and in Table 6.

Doubling time.
This is the most meaningful growth parameter when the history and development of a scientific field are under consideration. Refer to eqs. (6) and (7) for exponential growth, and to eq. (13) for logistic growth. As shown in Table 6, doubling times for the two growth phases for each of the Mars and the Magnetic Stratigraphy data estimated by the GEOREF computerized bibliographic database (dN/dt) and from citation analysis through the datestack ($S_1(D)$) lie in the rapid growth range. The 8 growth rates average to a doubling time of 4.2 yr.

RATE OF SELECTION

Equation (30), using smoothed values of $N'(D)$ and $S_1(D)$, was used to calculate the averaged rate of selection of references r by authors in the Mars and Magnetic Stratigraphy collections of source papers. These rates are shown in Table 6. The results are as follows.

Time range for selection rates.

$\bar{r}$ is averaged over all of the source papers used to determine $S_1(D)$. If the useful range of ages in citation date analysis is 2-10 yrs, then $D+2 < S < D+10$ for the group of source papers defining $\bar{r}$. Thus the latter is an average over a span of 8 years.

Geology and Geophysics of Mars.
The calculated values of $\bar{r}$ are nearly constant from 1966 to 1975. Each value is an average for a span of 8 years. The average value is

$$0.010 \pm 0.002 \text{ (standard deviation)}, \qquad (31)$$

which equals a choice of one reference for each year out of 100 published references in the field. The rate cannot be calculated beyond 1975 because of lack of sufficient width in the Datestack window. Thus $\bar{r}$ for the Mars data is defined only in Phase 1 of the growth of the field.

Magnetic Stratigraphy.
The rate of selection is divided into three periods. For the first period (1960-65) the rate is constant and given by :

$$r(D) = 0.092 \pm 0.020 \qquad (32)$$

which approximately equals an average choice each year of 1 out of 11 of the published references in the field. This period lies in Phase 1 of the growth of the field. It is followed by a transition period (1966-1971) in which $\bar{r}(D)$ drops through an almost linear transition. It reaches a constant value (1972-74) of :

$$\bar{r} = 0.017 \pm 0.001 \qquad (33)$$

This value approximately equals an average choice each year of one out of 59 of the published references in the field and is defined in Phase 2 of the growth. The values in the transition period are probably artifacts of the 8 year averaging timespan. For the Mars data both the computerized bibliographic database (GEOREF) and the Datestack methods indicate that the earlier phase grow more slowly (4.7 yrs, average of the two doubling times) than the later phase (t_d = 3.6 yr, average). The Magnetic Stratigraphy data do not indicate a similar pattern.

ESTIMATING GROWTH RATES FOR A LOGISTIC FROM CITATION ANALYSIS

Since from eq. (9) $y'/y = \alpha - \frac{\alpha}{K} y$, then if we have a model in which $\bar{r}$ is constant through a logistic growth phase, then

$$y' = \bar{r}\, S_1(D) \tag{34}$$

and if y is the cumulative of y',

$$y = \bar{r}\, \Sigma\, S_1(D) \tag{35}$$

Thus the value of $\frac{S_1(D)}{\Sigma S_1(D)}$ for $\Sigma S_1(D) = 0$ will give α, even if we do not know what $\bar{r}$ is. Then from eq.(13), doubling time can be calculated for any growth phase with constant r.

JOHN MICHELL ON THE CAUSE OF EARTHQUAKES

One potentially important application of citation analysis is to the history of science, to estimate growth curves for early periods in which no reliable data on the growth of science exist. In 1761, John Michell wrote a defining paper in the field of seismology [13]. The practice of formally listing citations at the end of papers had not begun, but Michell included a thorough set of references to earlier published works in the field.

There is an unusual opportunity in this case to test Michell's rate of sampling the published literature using an excellent later bibliography. During the period 1855 to 1865, Alexis Perrey published his personal bibliography of seismology, the field in which Michell's paper falls [14]. Perrey was a leading figure in seismology, a prolific writer, recognized as a pioneer. His bibliography is extremely thorough, and Michell's citations lie within the period covered by the bibliography. The cumulative total of these citations is given in Table 7. On Fig.10, the cumulative values from Table 6 multiplied by 2.7 is shown. The figure shows these multiplied values superimposed with the portion of Perrey's bibliography [14] (Fig.9) coinciding with the time span of Michell's [13] citations. We see that Michell's rate of selection of literature was constant over the period 1600-1750. His rate was r = 0.37, indicating that he cited one out of approximately every 3 published bibliographic items.

This is an unusual case in which the citation numbers of a single paper have a large signal/noise ratio. However, it is a review paper as well as a pioneering definition of the science of seismology. Such papers often are reliable in their citation frequencies. An important future development is to investigate the statistics of stacking for small numbers of source papers so that signal/noice can be optimized in order to estimate the growth of fields in the early history of modern science. For this purpose robust statistical methods for fitting models to noisy data in the presence of non-Gaussian noise should be investigated.

CONCLUSIONS

1. That the number of citations of a given age are jointly proportional to the author's rate of selection and the pool of literature (N'(D) approximately) available in the year corresponding to that particular age is a premise accepted for the past half century in Scientometrics. Results in the present paper confirm this basic principle.

2. The rate of selection is a diagnostic quantity, varying from period to period in science. It is one of the indicators of the general scientific climate within each period of the development of a given specialty.

3. Stacking is a simple and efficient way to combine citation data from a collection of source papers; stacking can be done on the basis of dates of citations, or on the basis of ages of citations. Datestacking is preferable because the rate of selection of citations can be handled in a more direct way than it can be in stacking by age.

4. The basic equations governing date and age stacking are developed for the general case (any growth model).

5. Three collections of source papers from fields in geoscience : the Geology and Geophysics of Mars, Magnetic Stratigraphy, and early Seismology were used for a study of growth of science literature. The first two fields are rapidly growing sciences, and the third less so for the period in which it was studied. Growth of literature in a field was taken as an indicator of the growth of the field as a whole.

6. The Mars collection led to stacked citation data covering the period 1962-1976; the yearly accumulation of literature (N'(D)) was also estimated from GEOREF, a comprehensive computerized bibliographic database for geoscience, for the period 1962-1985.

7. The data for Mars are best fit by two logistic phases, except that the time window for stacking allows for only an exponential estimate of growth for the second phase (Table 6). Doubling times average 4.2 yr.

8. The Magnetic Stratigraphy collection led to stacked citation data covering the period 1958-1977; and the yearly accumulation of literature (N'(D)) was estimated from GEOREF for the period 1962-1977.

9. The data for Magnetic Stratigraphy are best fit by two logistic phases of growth (Table 6). Doubling times average to t = 4.2 yr.

10. Early Seismology (John Michell) was found to be growing with doubling time about 25 yr, in the approximately exponential early part of a logistic growth phase.

11. Rates of selection are interesting indicators which might help in gauging the general conditions which affect various periods in science. The combination of N'(D) data from the GEOREF database with citation analysis leads to calculation of this indicator.

12. Selection rates were calculated and have the following values :
    (a) Mars collection, 1966-1975, 0.010 ($\approx$ 1/100);
    (b) Magnetic Stratigraphy, 1960-1965, 0.092 ($\approx$ 1/11), and 1972-74, 0.017 ($\approx$ 1/59);
    (c) Early seismology (John Michell), 1600-1750, 0.37 ($\approx$ 1/3).

13. Constant rates for periods of time are apparently common, and the patterns are different for different sciences and different periods of time.

14. Data on the rate of selection should be extended to many fields and through a range of time. Methods of theoretical prediction of values of r(D) through a span of sciences and time should be developed. Values of r(D) then allow date stacks to be used to estimate N'(D), thus extending our quantitative knowledge of growth in science.

ACKNOWLEDGMENTS

This research and its presentation at the Second International Conference on Bibliometrics, Scientometrics and Informetrics was supported by research grant no. 410-88-0763 from the Social Sciences and Humanities Research Council of Canada, and by a travel grant from the University of Manitoba.

Mrs. S.D. Fay carried out the searches of the GEOREF database to establish values of N'(D). She also carried out the calculations of the stacks and the plotting of the Figures, as well as the word processing of the manuscript.

Mrs. Esther B. Hall assisted in assembling data in the National Science Library (CISTI) in Ottawa and discovered a part of Alexis Perrey's Bibliography of Seismology which had been previously little known.

REFERENCES

[1] de Solla Price, J.D., Little Science, Big Science. Columbia University Press, 77-81, 1963.

[2] Moed, H.F. et al., "The Use of Bibliometric Data for the Measurement of University Research Performance". Research Policy 14, 131-149, 1985.

[3] Hall, D.H., "The Interface Between Geoscience and Industry : a Case Study of the Interaction between Research and the Discovery and Mining of Ores for Nuclear Fuels". Scientometrics 11, 1987.

[4] Reference 1, p.81.

[5] Menard, H.W., Science, Growth and Change. Cambridge, Mass., 1971.

[6] Hall, D.H., History of the Earth Sciences during the Scientific and Industrial Revolutions. Elsevier Scientific Publishing Co., The Netherlands, 19-25, 1976.

[7] Hall, D.H., "Rate of Growth of Science Literature in Geoscience from Computerized Databases". Scientometrics (accepted for publication November 15, 1988).

[8] Brown, R.J., G.H. Friesen, D.H. Hall and O.G. Stephenson, "Weighted Vertical Stacking in Crustal Seismic Reflections Studies on the Canadian Shield". Geophysical Prospecting 25, 1977.

[9] Davis, H.T., The theory of econometrics. The Principia Press, Bloomington, Indiana, 1941.

[10] Hewlett-Packard (HP-97) Standard Pac Manual; 03-01 and 03-04, 1976.

[11] Reference 6, p.23.

[12] Kennett, James P., Magnetic Statigraphy of Sediments. Benchmark Papers in Geology 48, 1979.

[13] Michell, John, "Conjectures Concerning the Cause and Observations upon the Phenomena of Earthquakes". Royal Society of London Philosophical Transactions 51(2), 1761.

[14] Perrey, M. Alexis, Bibliographie Seismique. Mémoires de l'Académie de Dijon, 2e serie, t.iv, 1-253 (1855); t.v, 87-192 (1856); t.x, 1-53 (1862); and t.xiii, 34-101 (1865).

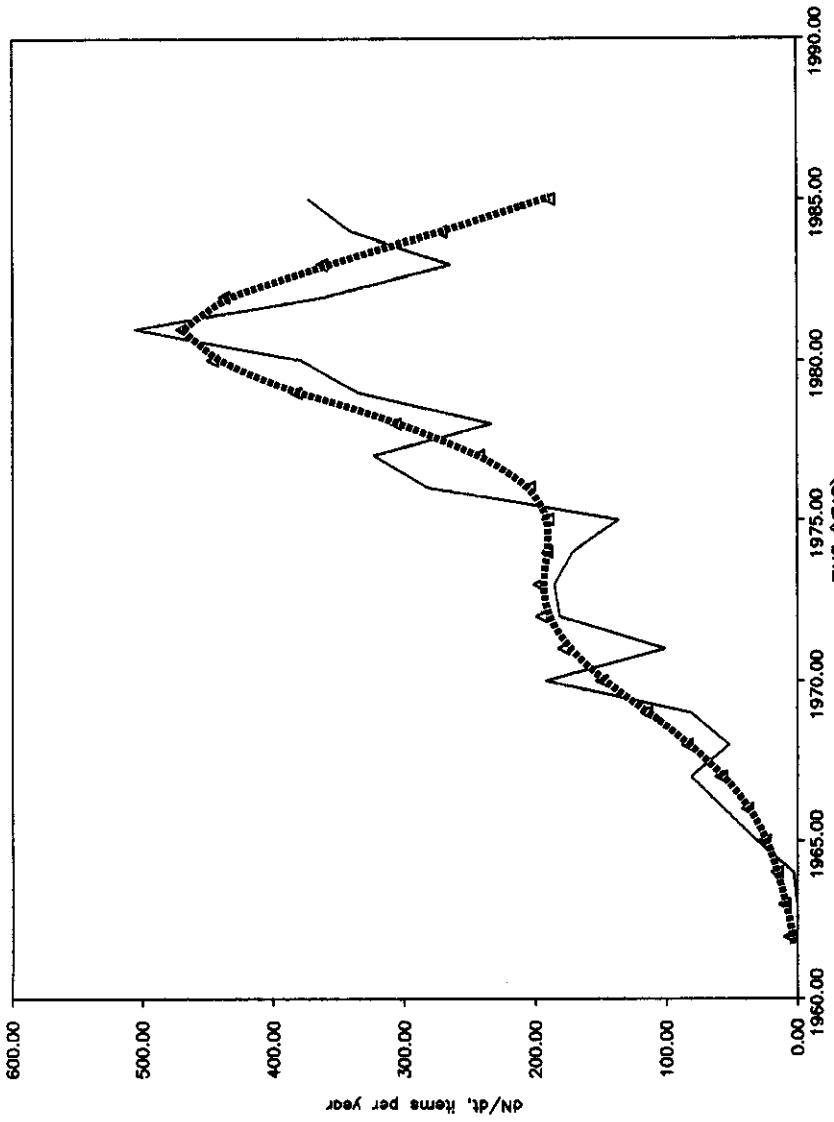

Fig.1 : (N'(D) items per year) 1962-1985, Geology and Geophysics of Mars, from the GEOREF computerized bibliographic database. Also shown are the best-fit overlapping pair of logistic curves (thick line and symbols). See Tables 4 and 6 and the text for details.

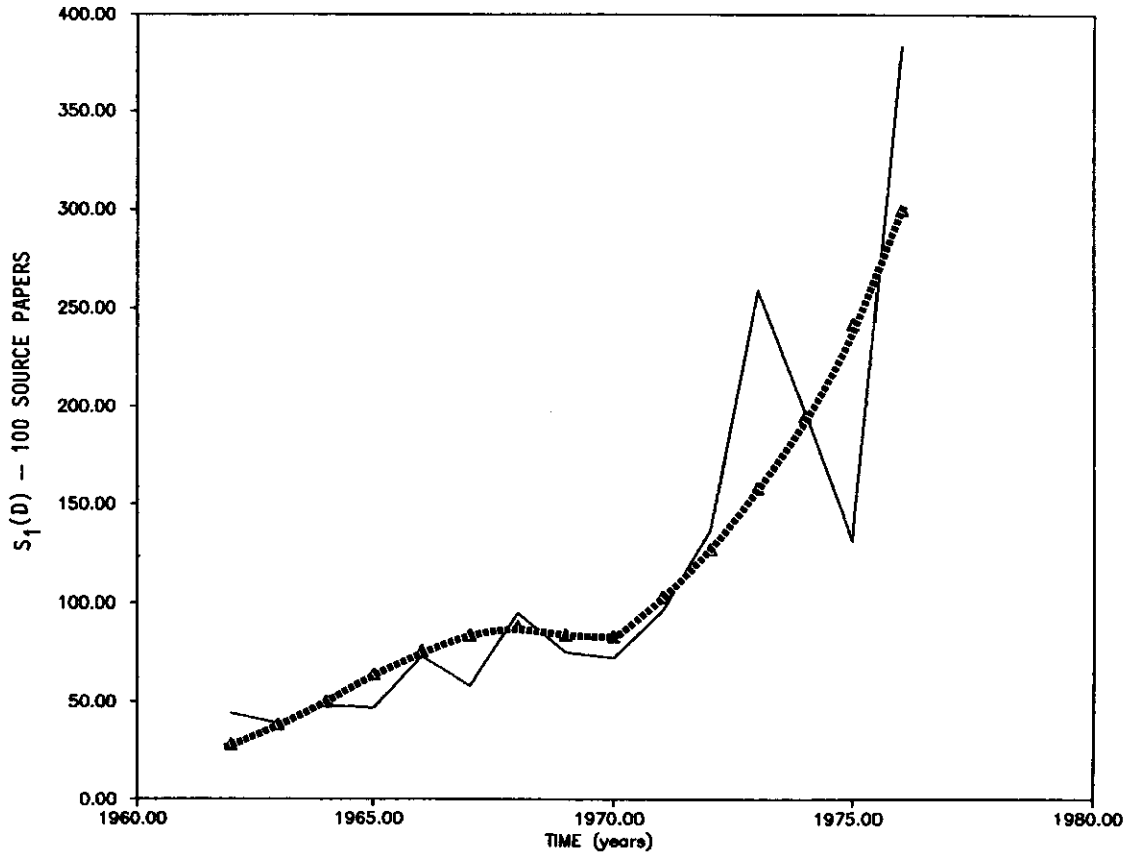

Fig.2 : $S_1(D)$, the Datestack, (items per year for 100 source papers), 1962-1976, Geology and Geophysics of Mars, from citation analysis. Also shown are the best-fitting first phase logistic followed by a non-overlapping exponential (thick line and symbols). See Tables 4 and 6 and the text for details.

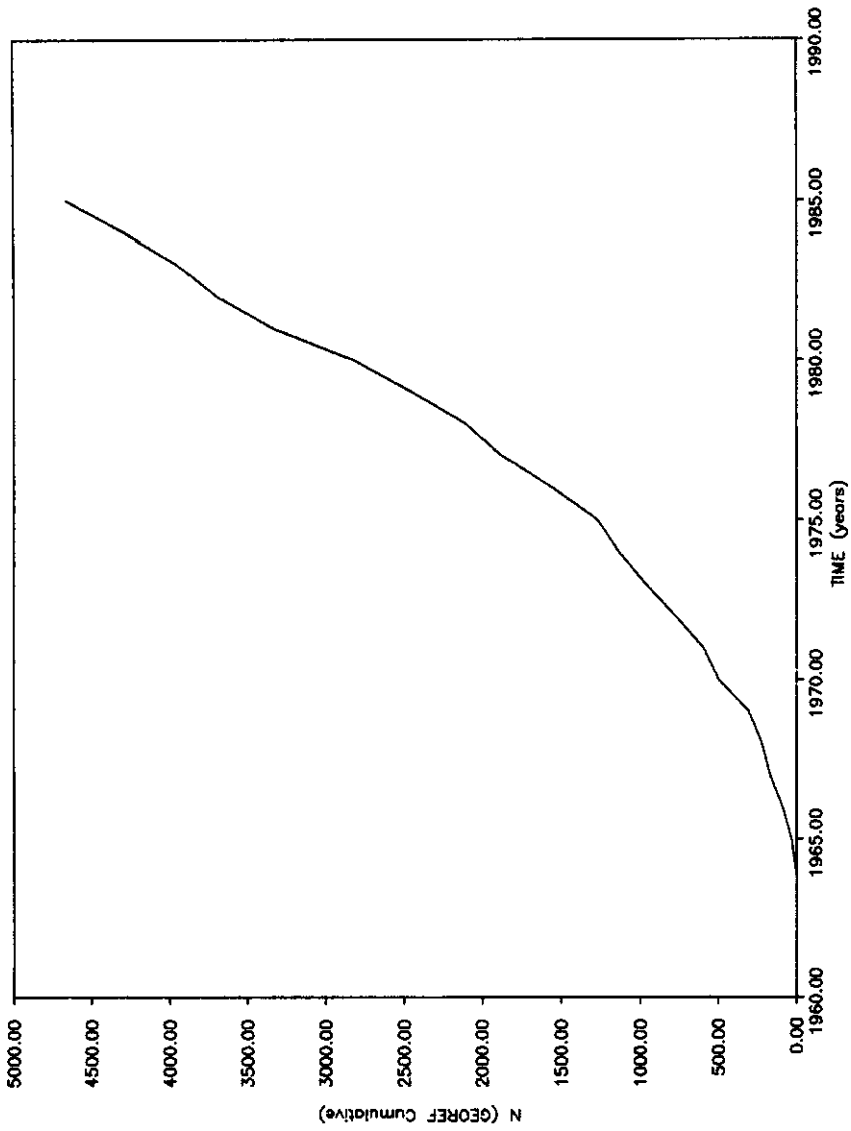

Fig.3 : Cumulative curve (N(D)) of the GEOREF data, (1962-1985).
Geology and Geophysics of Mars. See Table 4 and the text for details.

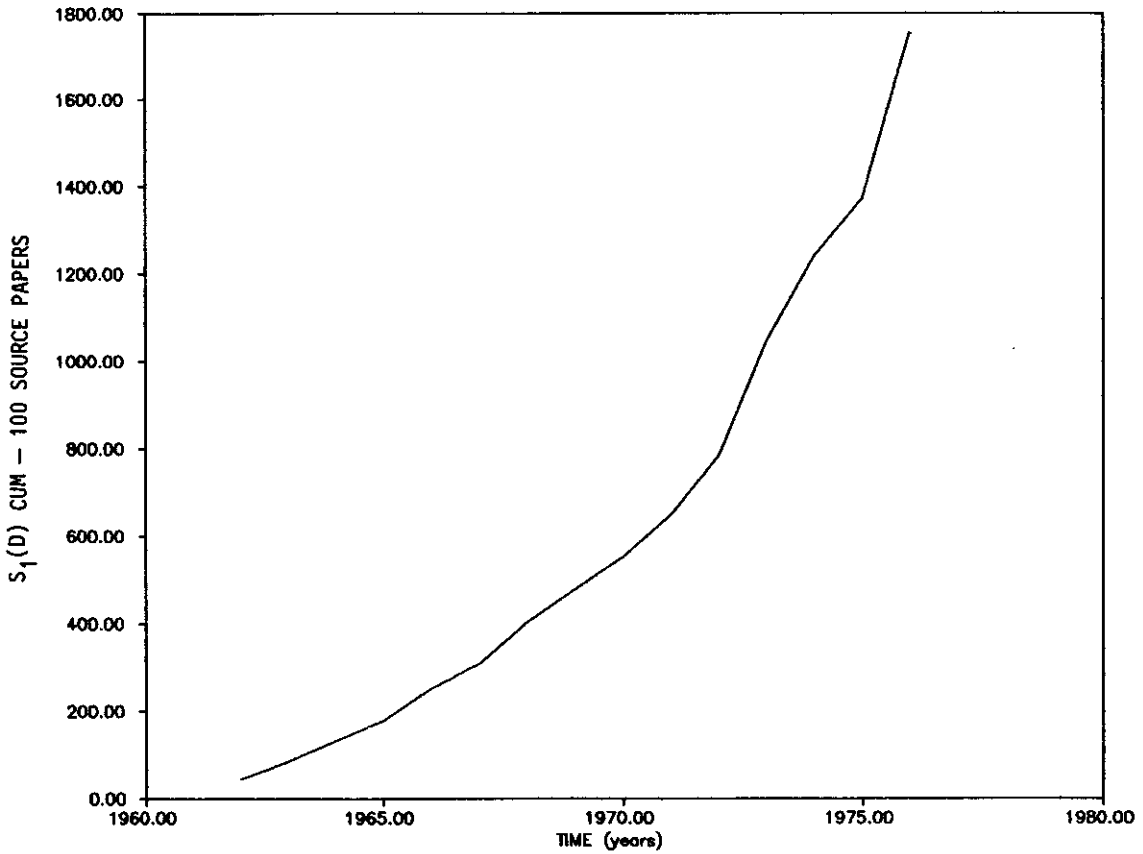

Fig.4 : Cumulative of $S_1(D)$ of the Datestack data (1962-1976).
Geology and Geophysics of Mars. See Table 4 and the text for details.

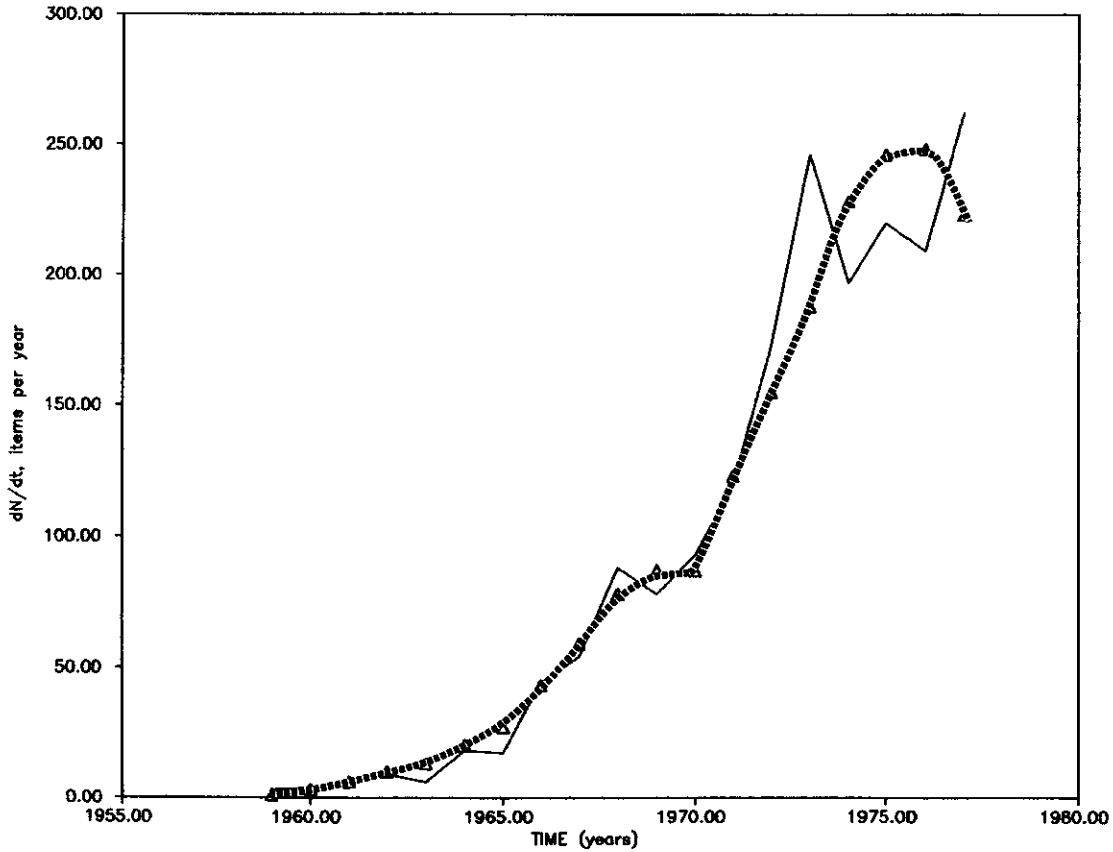

Fig.5 : N'(D) (items per year), 1959-1977, Magnetic stratigraphy, from the GEOREF computerized bibliographic database. Also shown are the best-fit overlapping logistic curves in escalation (thick lines and symbols). See Tables 5 and 6 and the text for details.

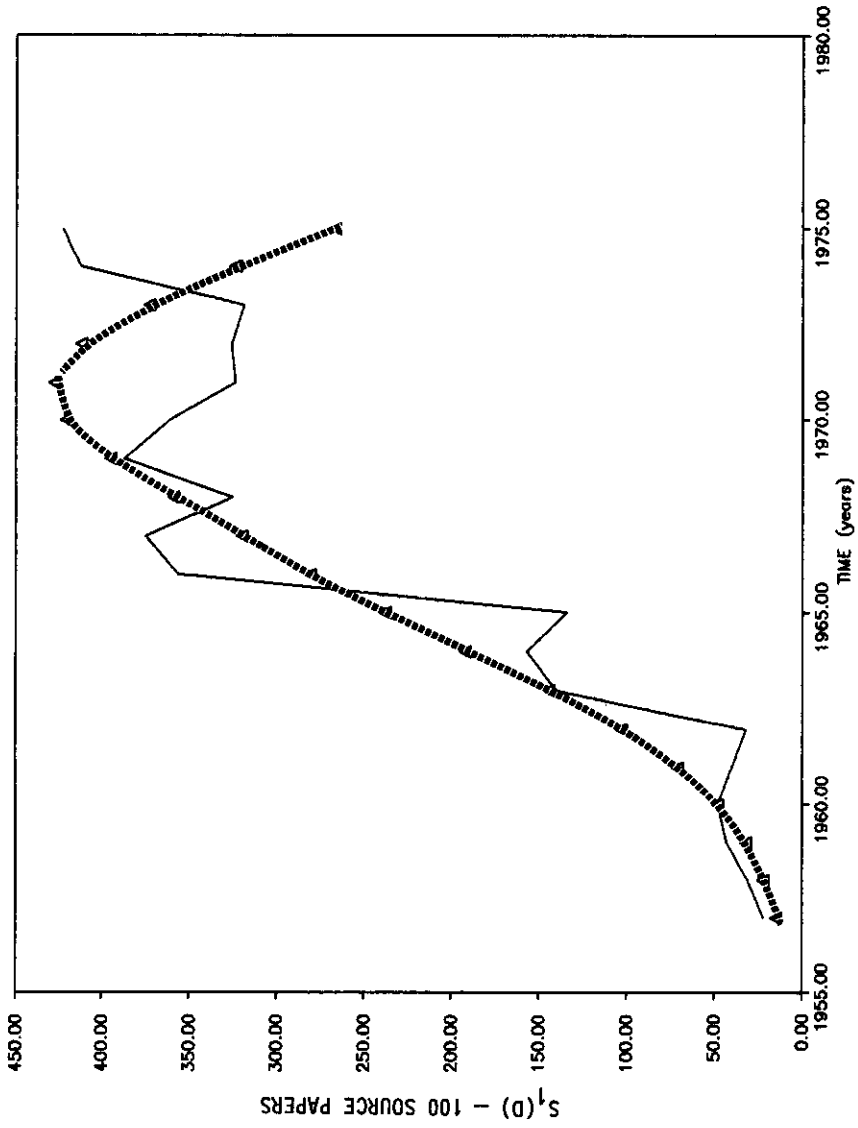

Fig.6 : $S_1(D)$, the Datestack, (items per year for 100 source papers), 1957-1975, Magnetic Stratigraphy. Also shown are the best-fit overlapping pair of logistic curves (thick line and symbols). See Tables 5 and 6 and the text for details.

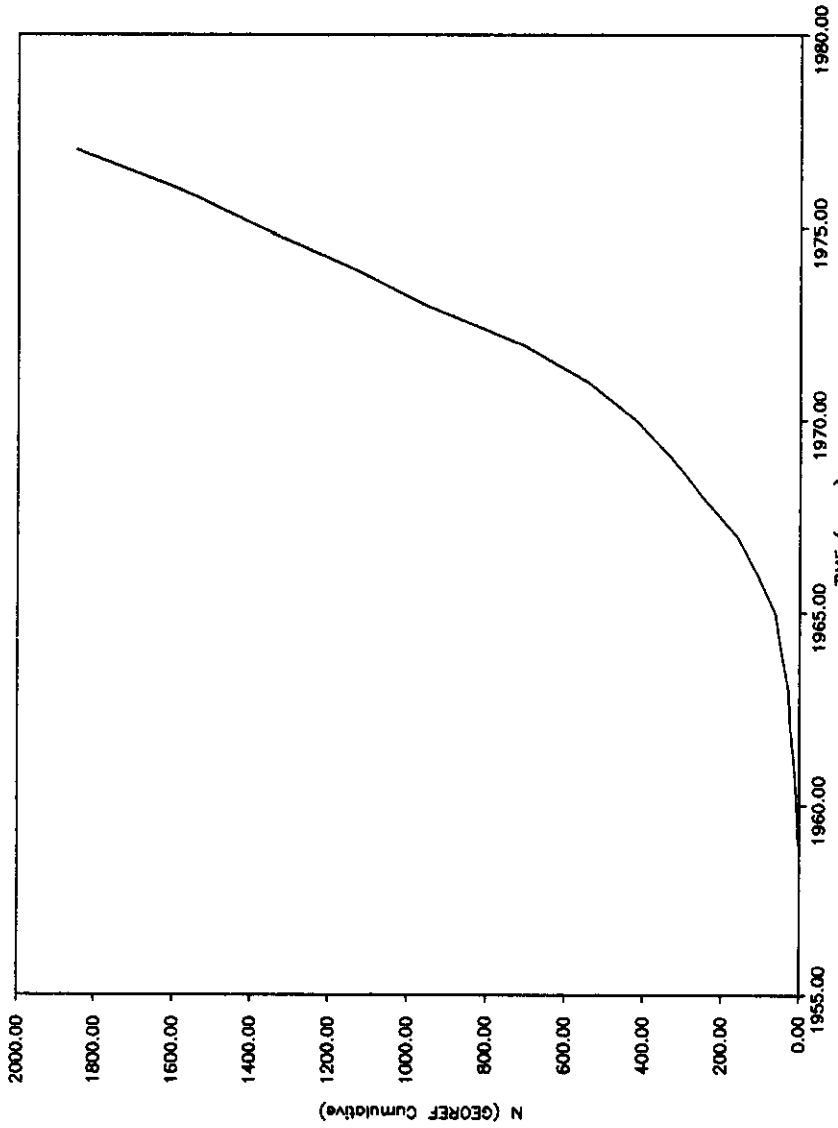

Fig.7 : Cumulative of (N(D)) for the GEOREF data, 1959-1977, Magnetic Stratigraphy. See Table 5 and the text for details.

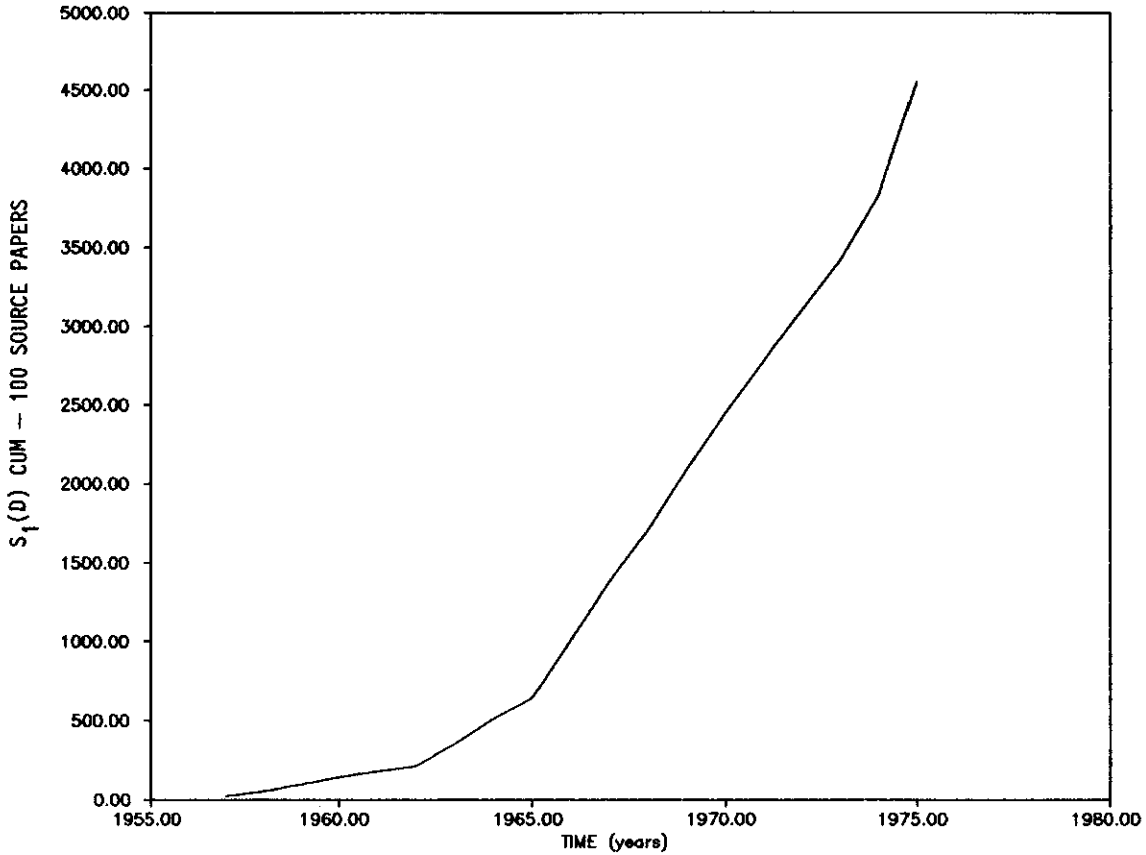

Fig.8 : Cumulative of $S_1(D)$, the Datestack data, 1957-1975, Magnetic Stratigraphy. See Table 5 and the text for details.

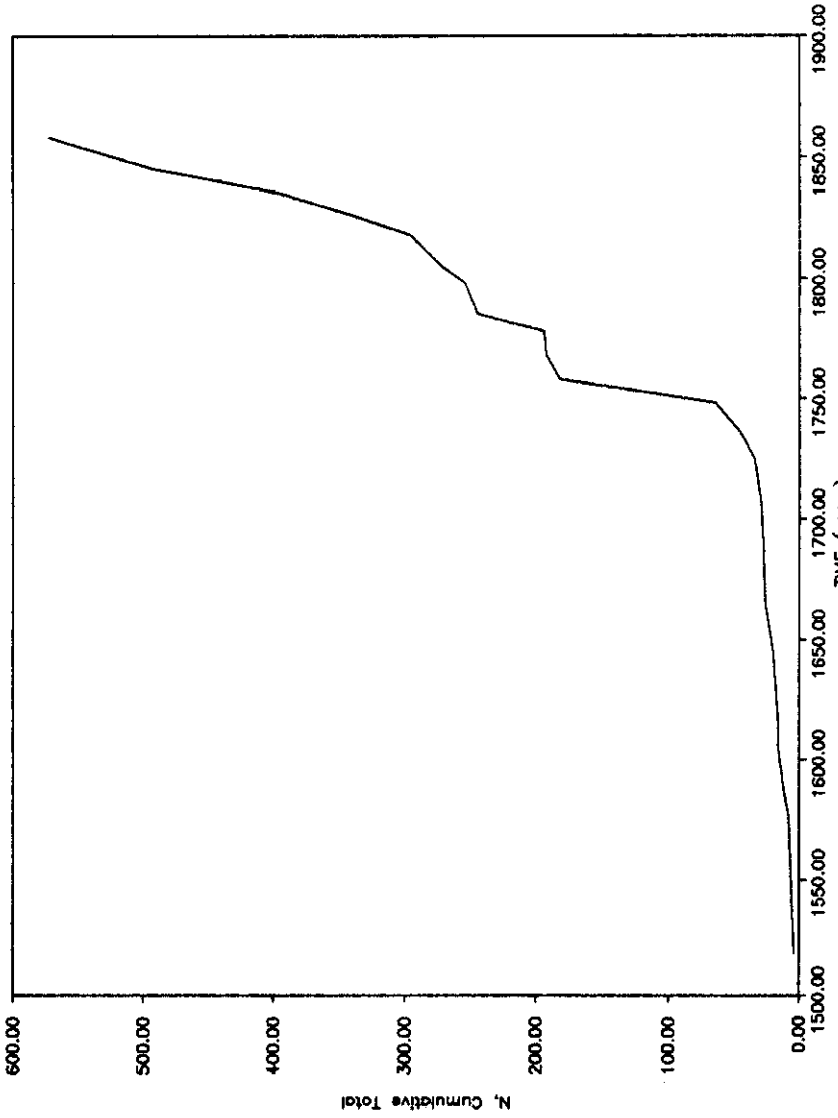

Fig.9 : Cumulative total of literature on seismology and the forerunners of that field, from Alexis Perrey's bibliography [13].

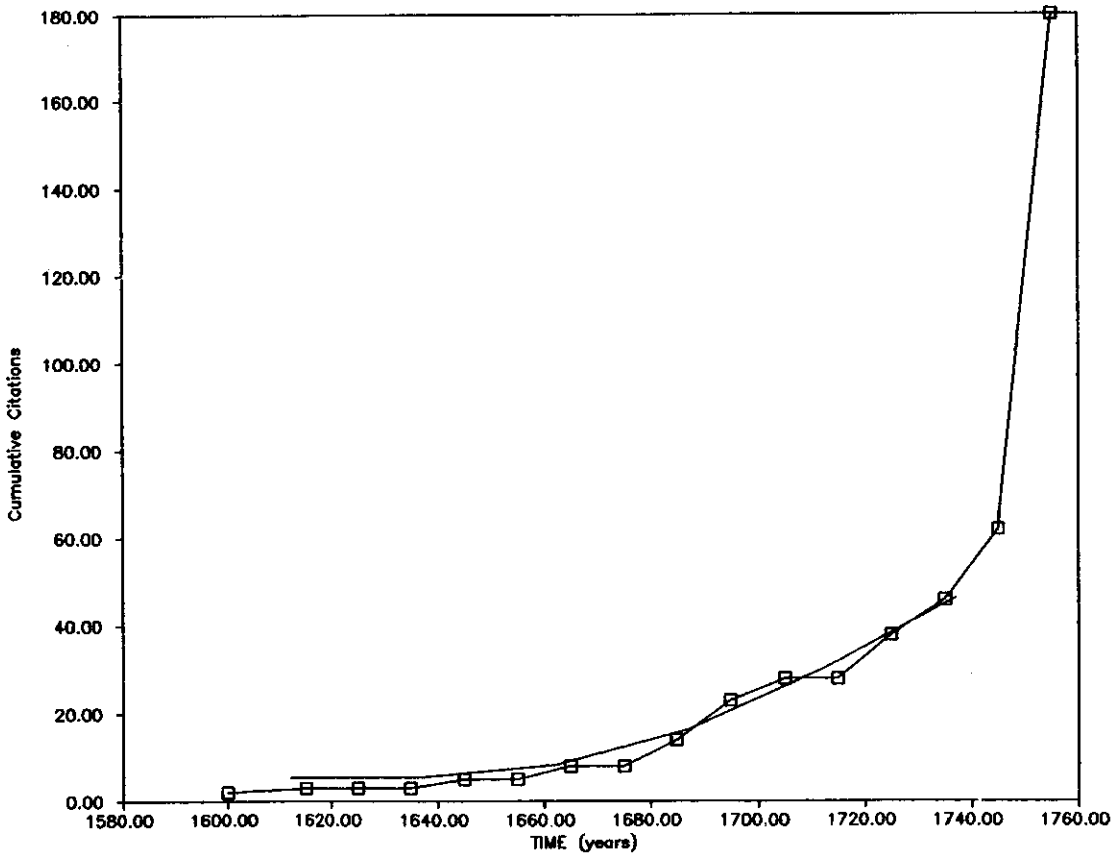

Fig.10 : Cumulative totals of Michell's citations (Table 6) in the period 1600-1750, multiplied by 2.7 and compared to Perrey's bibliography (Fig.9). This comparison shows that Michell cited at the constant rate of about 1/3.

Table 1

T(yr) / S(yr)	0	1	2	...		...
S_1	$C(S_1)_A$	$C(S_1 - 1)$	$C(S_1 - 2)$	...	$C(S_1 - T)$	...
S_2	$C(S_2)$	$C(S_2 - 1)$	$C(S_2 - 2)$	...	$C(S_2 - T)$	...
⋮	⋮	A' ⋮	⋮		⋮	
S_i	$C(S_i)$	$C(S_i - 1)$	$C(S_i - 2)$	...	$C(S_i - T)$	...
⋮	⋮	⋮	⋮		⋮	
S_L	$C(S_L)$	$C(S_L - 1)$	$C(S_L - 2)$	...	$C(S_L - T)$	...

AA' is a segment of a time line (see date stacking) assuming as an example that $S_2 - S_1 = 1$ yr.

Table 2* : Geology and Geophysics of Mars

Journal/Book Title	Date	Vol.	No. of Source Papers Used
Journal of Geophysical Research	1977	82	49
	1979	84	48
	1980	85	1
	1981	86	4
	1982	87	46
	1983	88	7
	1984	89	9
	1985	90	6
	1986	91	16
	1987	92	8
Geology of the Planet Mars, Benchmark Papers in Geology, v.54, 1980, Vivien Gornitz (editor)	1965-66 1969 1971-79		35
TOTAL			229

*Summarized from work sheets filed in our laboratory.

Table 3 : Examples of citation data and date stack paths through the data

Source Paper No.	Reference : Geology of the Planet Mars Gornitz, V. (ed.) Benchmark Papers in Geology, v.48 Dowden, Hutchinson & Ross. 414 p.	Age (Yr)																	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1977 : 29	p.361	1	10	1	-	-	3	1	1	-	-	-	-	-	-	-	-	-	-
30	p.365	-	9	2	1	-	2	1	1	-	-	-	1	-	1	1	1	1	-
31	p.379	3	3	-	3	2	-	1	-	-	-	-	-	-	-	-	-	-	-
32	pp.278-280	2	11	5	10	5	5	3	-	4	1	3	2	-	1	-	2	1	-
33	pp.331-222	8	20	1	6	9	1	2	-	-	1	-	-	-	1	-	2	-	1
34	pp.328-329	3	10	-	4	1	3	2	3	1	1	-	1	-	3	1	-	-	1
1979 : 35	pp.381-395	-	7	38	43	18	25	37	24	12	4	16	4	4	11	10	5	1	2
					1976						1970								1962

Table 4 : Mars GEOREF Database and Datestack Values

Date	dN/dt (GEOREF)*	Cumulative; N(GEOREF)*	$S_1(D)$ Datestack#	Cumulative Datestack#
1962	0	0	0.44	0.44
1963	0	0	0.39	0.83
1964	3	3	0.48	1.31
1965	31	34	0.47	1.78
1966	56	90	0.73	2.51
1967	81	171	0.58	3.09
1968	52	223	0.95	4.04
1969	81	304	0.75	4.79
1970	192	496	0.72	5.51
1971	101	597	0.96	6.48
1972	181	778	1.37	7.85
1973	185	963	2.59	10.44
1974	171	1134	1.97	12.41
1975	136	1270	1.31	13.72
1976	282	1552	3.84	17.56
1977	323	1875		
1978	233	2108		
1979	334	2442		
1980	378	2820		
1981	505	3325		
1982	360	3685		
1983	264	3949		
1984	339	4288		
1985	372	4660		

* all represent items/yr;

datestacks are per source paper

Table 5 : Magnetic Stratigraphy : GEOREF Database and Datestack Values

Date	dN/dt (GEOREF)*	Cumulative; N(GEOREF)*	$S_1(D)$ Datestack#	Cumulative Datestack#
1957			0.22	0.22
1958			0.31	0.53
1959	3	3	0.43	0.96
1960	4	7	0.48	1.44
1961	6	13	0.4	1.80
1962	9	22	0.32	2.12
1963	6	28	1.4	3.52
1964	18	46	1.57	5.09
1965	17	63	1.34	6.43
1966	44	107	3.56	9.99
1967	54	161	3.75	13.74
1968	88	249	3.25	16.99
1969	78	327	3.87	20.85
1970	93	420	3.61	24.47
1971	120	540	3.24	27.71
1972	173	713	3.26	30.97
1973	246	959	3.19	34.16
1974	197	1156	4.12	38.28
1975	220	1376	4.23	45.51
1976	209	1585		
1977	262	1847		

* all represent items/yr;

datestacks are per source paper

Table 6 : Parameters of Growth Phase

MARS			
GEOREF (dN/dt)			
PHASE 1 (logistic)			PHASE 2 (logistic)
t_d (doubling time)	4.3 yr		3.9 yr
dominant	1962-1975		1975-1985
alternative	1962-1973[2]	Table 4 & Fig.1	1973-1985[2]
Datestack $S_1(D)$			
PHASE 1 (logistic)			PHASE 2 (exponential)[1]
t_d	5.1 yr		3.3 yr
dominant	1962-1970	Table 4 & Fig.2	1970-1976
Rate of Selection			
PHASE 1			PHASE 2
r(D) = 0.010	0.002	1/100 (1966-1975)[3]	Undefined[3]
MAGNETIC STRATIGRAPHY			
GEOREF (dN/dt)			
PHASE 1 (logistic)			PHASE 2 (logistic)
t_d	4.3 yr		3.5 yr
dominant	1962-1970		1970-1977
alternative	1962-1968[2]	Table 5 & Fig.5	1968-1977[2]
Datestack $S_1(D)$			
PHASE 1 (logistic)			PHASE 2 (logistic)
t_d	3.4 yr		5.7 yr
dominant	1957-1965	Table 5 & Fig.6	1965-1977
Rate of Selection			
PHASE 1		TRANSITION	PHASE 2
r(D) = 0.092 (defined 1960-1965)	0.020 1/11	a linear drop in r(D) (1966-1971)	r(D) = 0.017 0.001 1/59 (1972-1974)

Notes by Table 6 :

[1] See text for discussion.

[2] This would give an allowable but less closely fit logistic.

[3] The period through which r(D) can be defined is limited by the time window for datestacking (1965-1976) and after smoothing 1966-1975).

Table 7 : Cumulative total of Michell's [12] citations

Dates	1600-25	1625-50	1650-75	1675-1700	1700-25	1725-50
Cumulative total	2	2	3	6	11	17