Numerical Methods of Bibliographic Analysis

B. C. BROOKES

IT IS ONLY in the last eight or ten years that the numerical aspects of bibliography have attracted attention, although some of the numerical regularities that occur in bibliography have been known for thirty or forty years. Results are, therefore, still meager and applications are still few. Moreover, most of the work so far reported has been limited to numerical analysis of the literature of the natural sciences. This is in part because the secondary sources in the natural sciences are the best organized and so provide the most accessible data; in part because the literature of the natural sciences are the least restricted by linguistic barriers; and in part because the proposed world-wide systems, such as those advocated in the UNISIST report, offer an immediate field of application in the design of economic and efficient systems based on the results of numerical bibliographic analysis. However, the field of possible application is gradually widening: serious efforts are now being made to organize the more diffused literatures of the social sciences, for example.

The main practical purposes that numerical analysis can serve are based on the belief that quantification is a necessary component of the design of economic information systems and that measurement of the key processes of an information system is a necessary component of management control. The present main objectives of numerical analysis are:

1. the design of more economic information systems and networks;
2. the improvement of the efficiencies of information-handling processes;
3. the identification and measurement of deficiencies in present bibliographical services;
4. the prediction of publishing trends; and
5. the discovery and elucidation of empirical laws which could form the basis for developing a theory of information science.

B. C. Brookes is Reader in Information Science in the School of Library Archives and Information Studies, University College of London, England.

Estimating the Completeness of a Bibliography

Strictly speaking, no bibliography can ever be *proved* to be complete. The bibliographer can only publish his final list and so, in effect, challenge all who may be interested to point to any omission they may note. The bibliographer who strives to prepare a comprehensive bibliography will usually turn first to the most accessible and productive sources contributing to his subject. But as his work progresses he has to seek and identify other less productive sources. The search, ever more penetrating and wide-ranging, is continued while relevant items are found. But new finds occur with steadily decreasing frequency in spite of continued effort. There is no positive signal which indicates to the bibliographer when his search has been completed; the only sign is the absence of further relevant items.

However, the law first formulated by S. C. Bradford,[1] and known by his name, offers the possibility of estimating the number of sources and the number of items that one can expect to find. This estimate is based only on knowledge of a small but sufficient number of the most productive sources. Unfortunately, Bradford formulated his "law of scatter" in two versions which Wilkinson[2] has recently shown to be formally different although closely similar. But both formulations lend themselves to methods of estimating the size of a comprehensive bibliography if the subject and the range in time are first well-defined.

In its original form, Bradford's law said nothing about comprehensiveness. But obviously a bibliography must be finite and the number of items produced by the least productive sources cannot be less than one. When this consideration is expressed in one of the two formulations of Bradford's law, a simple graphical technique for estimating the size of the complete bibliography can be devised. It requires the drawing of the "bibliograph."[3]

The most productive sources are first ranked in decreasing order of productivity. The cumulative sums of items found from these sources are then plotted on a graph as shown in figure 1. The most convenient graph paper for this purpose is semilogarithmic; the linear scale is applied to the cumulative sums and the logarithmic scale (which, although marked 1, 2, 3 . . . 10, 11, 12 . . . in successive digits, actually spaces them according to the logarithms of the digits) is used to indicate the ranks of the sources. So, on the graph, point A indicates the number of items yielded by the most productive source, point B indicates the number of items yielded by the two most productive sources together, and so on. The first few points will be found to lie on a rising curve which, sooner

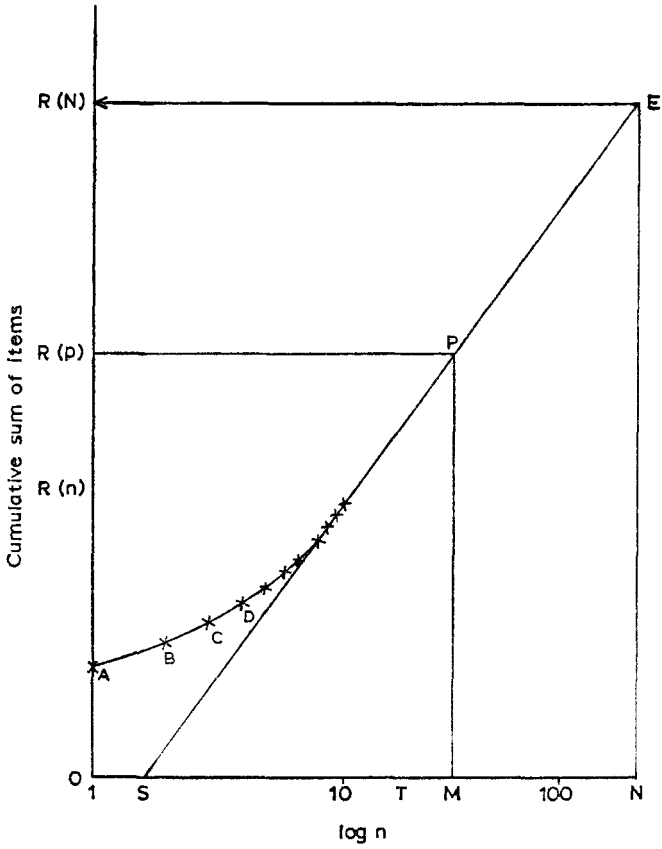

FIG. 1. Plotting the Bibliograph

or later, will run into a straight line. As soon as the straight line is definite enough, an estimate of the end-point of the line can be made.

The straight line is extended to meet the axis of log n (at S in figure 1) and some convenient point P is marked on the straight line. It can then be shown that the total number of sources, N, expected to contribute to the bibliography is given by

$$N = \frac{R(p) \cdot OT}{3\,SM} \tag{1}$$

where $R(p)$ is the number of items corresponding to the point P and where the lengths OT and SN are measured as accurately as possible in millimeters.

When N is known, it is possible to estimate the total number of items to be expected. It may be possible to mark the point N on the graph and so to mark the corresponding point E on the continuation of straight line SP. The required value of $R(N)$ can then be read from the scale of $R(n)$ on the left hand side. Alternatively, $R(N)$ can be calculated from the formula

$$R(N) = N \log_e N/S \qquad (2)$$

where it is the number corresponding to S on the log n scale. (A table of "natural" logarithms is needed.)

How realistic is the estimate? It is not possible to *prove* that E *must* be the end-point of the line because there is no logical reason why a comprehensive bibliography should conform so precisely to a mathematical law. Yet the technique has now been tested many times, especially against computer-produced bibliographies derived from retrospective searches of large data bases such as MEDLARS, and seems to be realistic. The advantage of using computer-produced bibliographies of this kind is that the items found must all conform to the search question as formulated for the computer search program. The relevance to the subject specified by the question is therefore uniformly controlled and it does not matter (for the purpose of testing the technique) whether the search question is "correctly" formulated or not. In every such case real and predicted end-points are very close indeed.

When the technique is applied to manually produced bibliographies the real and the estimated end-points can differ appreciably. A common occurrence is illustrated in figure 2. Here, if the plotting of the graph is continued beyond the points required to determine the straight line, the plotted points may fail to maintain the linear climb and fall away in a droop to end at some point such as G. In such cases it is plausible to argue that the bibliography is not complete in the sense in which the technique requires. For example, the bibliographer may quite reasonably state that he has been selective. He may say that he has noted only those items which are "of professional interest" or that he has omitted exact translations of some items into other languages. But, in our own experience, wherever a droop has been observed, it has always been possible to indicate either some selectivity or some omissions.

It has been noted, however, that the point E as calculated slightly overestimates the total number of items, though not the total number of sources. This fault arises because, as the graph nears the end-point, the sources end with a number which provides three items, a larger number

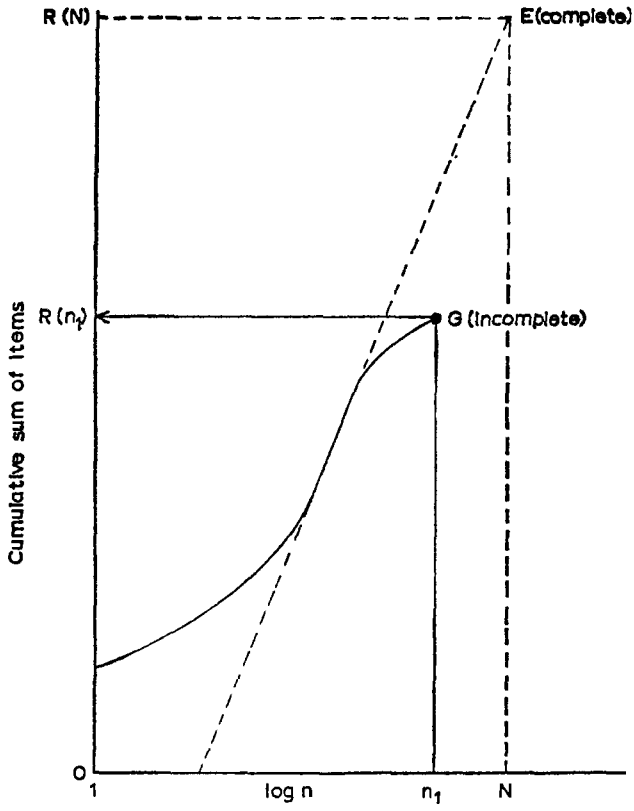

FIG. 2. The 'Droop': An Indication of Incompleteness

which provides two items and a still larger number of sources which provide only one item each. When the corresponding cumulative sums are plotted on a logarithmic scale they do not lie exactly on the straight line but form a series of lengthening arcs which intersect on the straight line (figure 3). The last one, however, is open at the end-point. When the complete data are carefully plotted, the graph ends in an open hook which ends slightly below the estimated end-point E. The "hook" however is clearly distinguishable from the "droop" mentioned earlier: the hook is concave upwards whereas the droop is concave downwards.

With these reservations Bradford's law can be used to provide reliable estimates—but of what precisely? The user of this technique has to appreciate that the end-point is determined by the items he includes in the data which initiate the curve and straight line and that these data

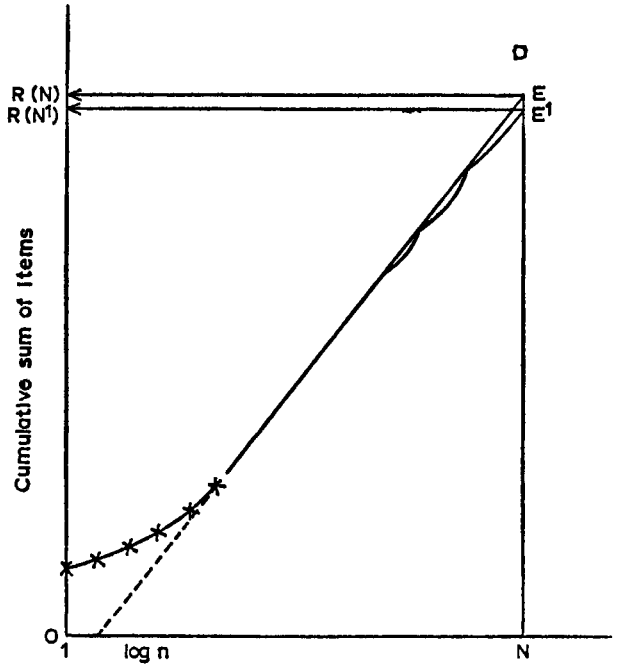

FIG. 3. The Arcs Near the End of the Bradford Linearity

define the totality he is estimating. For example, if it happens that these initial data are all derived from the most productive sources of, e.g., items published only in the *English* language, then the totality will be of items published in the English language. If the initiating data all refer to items published in, e.g., 1972 in any language, then the totality will be of items published in 1972 in any language. Most of the faults about the application of the technique that have been reported arise from lack of appreciation of the fact that the initiating data must be bibliographically well-defined in subject relevance and in period of publication and also be complete and exact as far as they need to go. And the reliability of the estimate is naturally critically dependent on the precision of the initiating data.

OTHER OCCURRENCES OF BRADFORD'S LAW OF SCATTER

The mechanism underlying the generation of a Bradford distribution seems to consist of two competing processes. The primary process is a

bandwagon effect: in bibliographical terms, authors would in general prefer to publish their papers in the core journals which correspond to the sources on the curved part of the bibliograph. But these journals receive more good papers on the particular subject than they can or wish to publish. So their standards of acceptance are high and this secondary restrictive process then pushes some papers out into the peripheral journals. The Bradford distribution follows: it represents an overall bandwagon effect which is modified by a restrictive effect over the sources at the core.

If this explanation were valid, the Bradford distribution should arise in other situations in which similar processes occur. It does. The items borrowed from a library follow the same law: the restriction on borrowing occurs because there are always a few items in such demand that some keen users have to wait.[4] The users too, ranked in order of the

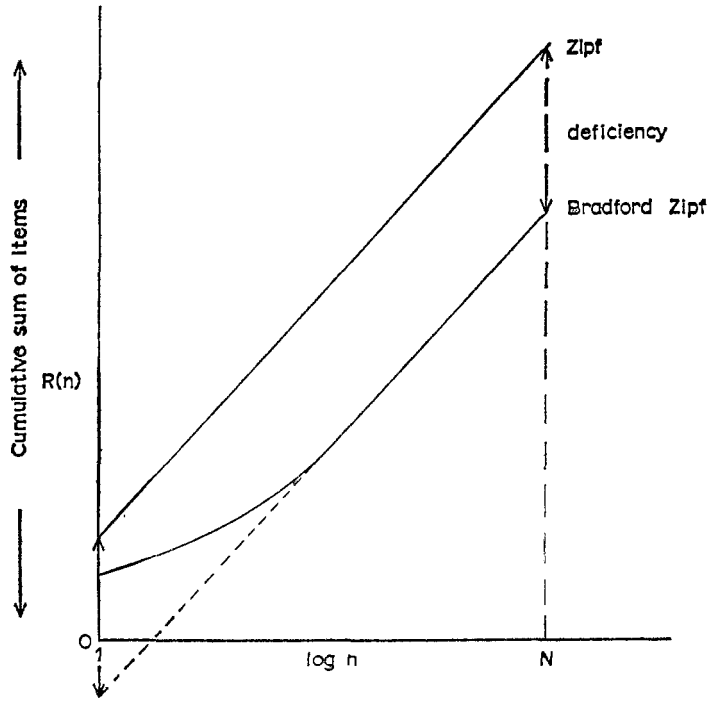

FIG. 4. Cumulative Zipf (Unrestricted) and Bradford-Zipf (Partially Restricted) Distribution

number of items they borrow, form another Bradford set: keen users compete with each other for the core items of the library and so are sometimes disappointed not to find the specific items they seek. It has been suggested (not too seriously) that a library could most economically be designed to provide the best possible service for all its users simply by striving to meet, as far as possible, the demands of its identified core users, because all other users will be content with what is so provided.

It has also been noticed that at a conference the contributors to the discussions, ranked according to the frequency of their questions or contributions, yield a Bradford set of data. Here, competition arises among the would-be contributors to "catch the chairman's eye," and the good chairman usually prefers to call a new contributor to the discussion when the chance arises. The index terms assigned to documents also follow a Bradford distribution because those terms most frequently assigned become less and less specific and therefore increasingly ineffective in retrieval. The publication of books by publishers follows the same law.

On the other hand, when there is no competition or other form of restriction, there is no core. At the present time, authors of scientific papers are generally allowed to cite as many references as they wish. The data which relate to items cited yield a bibliograph which consists of a straight line with no initial rising curve (figure 4). This is a cumulative Zipf distribution corresponding to the linear nonrestrictive part of the Bradford-Zipf distribution.

Applications of Bradford's Law

Comprehensiveness of abstracts and other bibliographic services

Every specialized interest in the natural sciences is now provided with abstracts services and some with retrospective search of cumulated data bases. These services are usually presented implicitly as though they covered their particular fields comprehensively, but it is well to check the coverage offered. The bibliograph can be revealing. If the graph ends in a droop it suggests that the service provided is not comprehensive, but before reaching such a decision the analyst must insure that his data are well-defined and are exact. For example, if it were decided to check the completeness of the list of items published in 1972, then adequate time must be allowed to permit the inclusion of all such items in the service being checked.

SELECTION OF SOURCES

The complete coverage of any well-defined subject is a costly objective, so most librarians are faced with the problem of deciding how to make optimum use of limited resources. If, for example, the attempt is made to identify the sources that contributed to the Bradford-estimated totality of some well-defined subject during the year 1970, it will be found that there may be 100 sources which provided only one item. If a similar analysis is made a year later, again it may be found that about 100 sources have yielded only one item each. But the two lists of one-item sources will not be identical: in each list there will be found sources which do not occur in the other. In short, whatever the subject may be, there is a very wide fringe of sources which contribute less frequently than one item per year, and continuous search is needed in this peripheral area to discover these occasional items. There is no discernible limit to comprehensiveness in this wide sense.

The librarian, therefore, has to be selective. The analysis of sources required to produce the data for the bibliograph requires the sources to be ranked in order of decreasing productivity. For a given annual expenditure on the acquisition of sources, it therefore provides the data needed to make selection a rational process, and the bibliograph enables the librarian to estimate the percentage coverage he can afford.

The core sources together always constitute a large fraction of the Bradford totality at relatively low cost and, because they also represent the most active sources, they can be regarded as providing the essential working minimum for any special library. Beyond this nuclear zone of the most productive sources there are sources which are gradually less and less productive. Moreover, there is some evidence (not yet fully established) that the nuclear zone produces the most frequently cited items and therefore, presumably, the most significant also. The general Bradford formula makes it possible to estimate how many of the most productive sources would yield any specified fraction p of the total number of items. The formula is

$$R(n) = N \log n/s \ (1 \leq n \leq N) \tag{3}$$

where

$R(n)$ = cumulative total of items contributed by the sources of rank 1 to n,
N = total number of contributing sources,
s = a constant characteristic of the literature.

Then

$$R(N) = N \log N/s \tag{4}$$

is the total number of items contributed by the N sources.

As a simple numerical example, consider a scientific literature which produces a total of 2000 items from 400 sources. We then have, substituting in (4),

$$2000 = 400 \log 400/s$$

so that

$$\begin{aligned} \log s &= \log 400 - 5.0 \\ &= 6.0 - 5.0 \quad \text{(natural logarithms)} \\ &= 1.0. \end{aligned}$$

It is now possible to complete the calculation, since

$$p = \frac{R(n)}{R(N)} = \frac{\log n/s}{\log N/s}$$

which on solving for $\log n$, yields

$$\log n = p \log N + (1 - p) \log s. \tag{5}$$

For the particular values of the numerical example, we have

$$\begin{aligned} \log n &= p \log 400 + (1 - p)\, 1.0 \\ &= 6.0\, p + (1 - p) \\ &= 5.0\, p + 1. \end{aligned}$$

If we now put $p = 0.8$, 0.6, 0.4 and 0.2 successively, the corresponding values of $\log n$ are 5.0, 4.0, 3.0 and 2.0 successively. And referring to tables of natural logarithms or exponentials, we find values of n as shown in the following table. Note in table 1 that the 20 most productive of the 400 sources contribute 40 percent of the total number of items.

TABLE 1

SOURCE COVERAGE BY PERCENTAGE

Percentage Coverage of Items	No. of Sources (n)	Percentage of Sources
100	400	100
80	148	37
60	55	14
40	20	5
20	8	2

The same estimates can also be found graphically. On semilogarithmic paper, with the rank n marked along the log scale and $R(n)$ along the linear scale, identify the two points: $R(N) = 2000$ with $N = 400$ and $R(n) = 0$ when $n = s = 2.7$. The line joining these points enables any corresponding values of n and p to be read directly from the graph.

These measures are, of course, purely quantitative; so it would be reasonable to consider sources on both sides of the proposed cut-off to insure, for example, that a source in a language unfamiliar to the users, or relatively costly, difficult to acquire or subject to delay is not retained in favor of another source, of slightly less productivity, in a familiar language, less costly or easier to acquire. The optimum cut-off has to be applied with judgment based on bibliographic knowledge of the subject area.

Any such cut-off leaves the librarian with the problem of searching for significant items provided by sources which he has decided not to acquire. Notice of such items should be given by the abstracts services or other secondary sources and their mode of acquisition will depend on the local circumstances. In Britain, the National Lending Library strives to acquire copies of all scientific sources and offers a photocopy by return mail of any item that is bibliographically well described. In such a case, the optimum cut-off point can be determined by estimating the cost, in terms of the delay to the users as well as of the actual monetary cost, and comparing this figure with the cost of acquiring items from the ranked sources as the rank (and therefore the cost per item acquired) runs from 1, 2, 3 . . . onwards until the two costs are equalized.[5]

This technique of determining the economic holdings of scientific libraries can be applied to any hierarchical system of libraries which consists of a number of small local libraries with special interests, a smaller number of regional libraries which service groups of special libraries, and a national library which offers the most comprehensive coverage. The conditions under which the regional libraries become redundant can be explored.[6]

The Measurement of Obsolescence

The world's knowledge of pure science is embodied in its published literature which is in part a description of theories and data that are wholly new, in part an updated version of earlier theories and a reinterpretation of older data in terms of the newer theories. So, at any given time, the current scientific literature is in part new and in part retrospective although the range of the retrospective view varies from one

science to another and even within a science, as in astronomy for example, there may be wide differences as there are between descriptive astronomy and astrophysics.

If a library's resources are limited and its users are primarily concerned with current research, there is little point in retaining in perpetuity all the back numbers of all the journals it takes. The usage of back numbers declines with age but the little-used volumes demand constant shelf space and maintenance services which add to the overall library costs. It is possible to estimate the optimum cut-off graphically when the aging rate is known, and economical to apply it when copies of any papers called for from the discarded volumes are readily available elsewhere.[6]

However, in measuring the rate of obsolescence of a literature we have to distinguish between the usage of the literature by those who contribute to it (who are its primary users) and the usage of the literature in any particular library. The first need is to measure the usage of the literature by those who contribute to it; the only publicly observable indicators of this usage are the references cited by the authors. The usage of a literature within any particular library will of course be of concern to the librarian, but his measures of usage will be of little interest to anyone else, nor are his data publicly observable. Citations within the literature to the literature, however, are publicly observable so that any results based on them can be checked by anyone who has access to the literature in any part of the world. If this publicly observable rate is known, then it should not be difficult to modify it to meet the usage patterns peculiar to any particular library.

The most straightforward technique of measuring obsolescence requires a sample of at least 2000 citations from the literature of the subject published within some specified year. The frequencies of citations to earlier papers *of the same literature* are then noted, taking the year as the unit. These frequencies are then cumulated in a table so that it is possible to read off how many citations were made *to each specified year or earlier*. The frequencies so cumulated, reducing year by year as the citations to still earlier years decline in frequency, are then plotted against their corresponding dates on semilogarithmic paper (two-cycle or three-cycle) with the frequencies marked along the logarithmic scale as in figure 5.

In general it will be found that after a hesitant start covering the two most recent years (because it takes that order of time to read a newly published paper critically, do the necessary scientific work, prepare the

paper and get it accepted and published), the plotted points fall steadily downwards on a straight line. This straight line continues until the period of the last world war is reached, when the plotted points usually move sideways (indicating an interruption of normal scientific publishing). Unless the sample is very large, the graph usually ends rather erratically as the frequencies become very small.

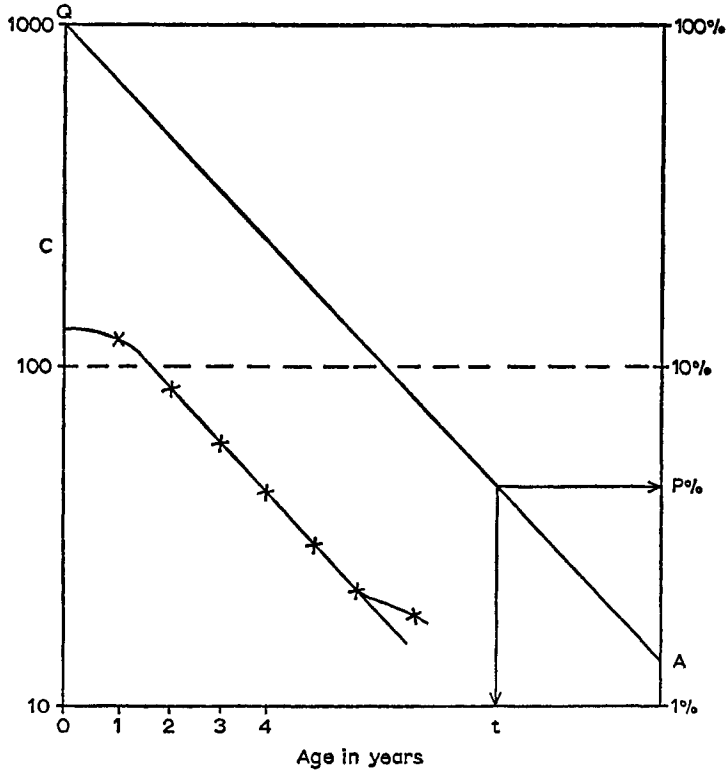

FIG. 5. Graphical Method of Estimating Obsolescence: C is the Number of Citations to the Year t or Earlier

The next step is to draw the best possible straight line through the *linear part* of the graph and to extend it to the edges of the graph paper. A second straight line QA is then drawn parallel to the first straight line and this second line becomes the means by which the annual aging rate and its effects can be read.

The theory of this procedure derives from the fact that the linearity

indicates that the citations as originally counted year by year can be expressed as the geometric sequence

$$R, Ra, Ra^2, Ra^3 \ldots Ra^{t-1} \ldots$$

where R is the presumed number of citations during the first year, some of which do not immediately emerge in publication. But as $a<1$, the sum of the sequence converges to the finite limit $R/(1-a)$. So the sequence plotted, which is the decreasing sum of citations to earlier years, can be expressed as

$$\begin{aligned} U_0 &= R(1 + a + a^2 + \ldots + a^{t-1} + \ldots) = R/(1-a) \ldots \qquad (6) \\ U_1 &= R(a + a^2 + a^3 + \ldots + a^{t-1} + \ldots) \\ &= Ra(1 + a + a^2 + \ldots + a^{t-2} + \ldots) \\ &= Ra/(1-a) = U_0 a \\ U_2 &= R(a^2 + a^3 + \ldots + a^{t-1} + \ldots) \\ &= Ra^2(1 + a + a^2 + \ldots + a^{5-3} + \ldots) \\ &= Ra^2/(1-a) = U_0 a^2 \end{aligned}$$

and so on. This result implies that the sequence plotted, i.e.,

$$U_0, U_0 a, U_0 a^2 \ldots U_0 a^{t-1} \ldots$$

is also a geometric series of the same ratio a as the original sequence. Hence, after t years the residual utility, $U_0 a^{t-1}$, is the fraction a^{t-1} of the original utility U_0. The value of this fraction can be read directly from the graph, either as a fraction or a percentage, as shown in figure 5.

Although the value of a can be read directly from the graph as the fraction corresponding to the age of one year, it is more accurate to read the fraction corresponding to the age eight years and then to take successive square roots. Thus if it is found that

$$a^8 = 0.36$$

then

$$a^4 = 0.60$$

and

$$a^2 = 0.775$$

giving

$$a = 0.88.$$

It is always helpful to draw the graph in case unexpected anomalies

arise which the librarian feels should be taken note of. Should any such anomaly make it difficult to draw any straight line with conviction, the constructed line OA cannot be drawn either. Any difficulty of calculation that arises, however, can be avoided by drawing a second graph beginning at Q which is geometrically parallel to the plotted graph. The required fractions or ages are then read by reference to the second graph. As the librarian can always exercise his judgment conservatively, there is no need for elaborate statistical procedures intended to yield figures of high precision.

If it has already been verified that, for the particular subject literature, the graphs can be assumed to be linear, it is possible to simplify the above procedure for estimating the numerical value of a. All that it is necessary to do is to divide the citation data into two categories when sampling and to count separately: (a) those eight years old or less and (b) those older than eight years. If these two counts amount to m and n respectively, then

$$m = R(1 + a + a^2 + \ldots + a^7)$$

and

$$\begin{aligned} n &= R(a^8 + a^9 + \ldots) \\ &= a^8R(1 + a + a^2 + \ldots). \end{aligned}$$

Hence

$$m + n = R(1 + a + a^2 + \ldots + a^7 + a^8 + a^9 + \ldots) = U_0$$

and

$$n = a^8U_0$$

so that

$$a^8 = \frac{n}{m + n}. \tag{7}$$

Square roots are then taken in succession as before to yield the required value of a.

It is emphasized that any value of a so obtained is the result of sampling and is therefore subject to sampling variance, i.e., two different samples of equal size drawn from the same literature are likely to give different results. Such differences decrease as the sample size increases but the sample size should always be stated when an obsolescence measure is determined.[7]

If the graph when plotted yields a slope which steadily decreases in

steepness, this result implies that the literature has at least two components of different rates of aging. A little knowledge of the subject usually makes it possible to separate the main components and to analyze them separately.

Obsolescence, Scatter and Growth

Librarians, concerned more with the usage of their own libraries than with the overall use of scientific literatures by those who create them, have been confused about the effects of the growth of literatures on their rates of obsolescence. They have argued that the value of a found by the technique described above has to be modified if the literature is growing. Their grounds for this argument rest on the idea that the probability of any particular paper being cited decreases as the number of papers published per annum increases. But they overlook the fact that scientific literatures grow, not because scientists increase their productivity, but because more scientists contribute to the literature at the same average rate. If this rate of growth of users of the literatures is taken into account, the probability of any new particular paper being cited remains constant. As far as earlier papers are concerned, the effect of growth is already implicitly included in the data collected. No corrections are needed.

In the usage of a *library,* however, other factors have to be considered. Thus, if, as in Britain, a new comprehensive library such as the NLLST is established, its usage will increase as its services become known and appreciated. It is possible that the rate of growth of usage of a scientific literature will equal or exceed the corresponding rate of obsolescence and the startling phenomenon of negative obsolescence may then be reported. It is, however, simpler to recognize that any literature ages at a uniform rate but that some libraries, especially new ones, can hope to attract usage at a rate which exceeds the rate of obsolescence. A simple analysis of the effect of growth on obsolescence is given in the Appendix.

There is some evidence, not yet wholly convincing, that scatter and obsolescence are positively correlated. It may be so, although at the present time there is no general agreement on how scatter should be defined or measured. The concept of scatter, except in the sense in which Bradford applied it to a comprehensive bibliography, is still vague. Yet there does seem to be a notion of scatter which is independent of completeness. For example, if 100 papers are selected at random from the literature of a subject and their sources are noted, one would expect that the number of sources so identified would be less than 100 and that the

fewer the number of sources, the less scattered the subject could be said to be. Unfortunately, the ratio of sources to number of papers selected at random can be shown to be a rather complicated function of the sample size. If the size of the sample of papers is increased from 100 to 200, the number of sources found in the sample of 200 is rather less than double the number found in the sample of 100. So, unless the sample size is standardized at, say, 100 papers, there is little prospect of finding a simple means of scatter by sampling. Although this problem is being studied, the related mathematics is surprisingly intractable.

It seems likely that scatter and obsolescence are related,[8] but that both are determined by rate of growth—the faster the rate of growth, the less is the scatter and the more rapid the obsolescence. A relationship between growth and scatter has been derived by Naranan[9] in mathematical terms, but bibliographic data are needed to test his theory in detail.

The literatures of most scientific subjects continue to grow exponentially, as judged by the number of papers published per annum, with a rate of growth over all subjects which has doubled the annual output in approximately ten years. The new literature makes its appearance dramatically in the copies of journals which a library receives every working day. The effects of obsolescence are much less dramatic: there is no external sign to indicate when the scientific utility of a journal has faded

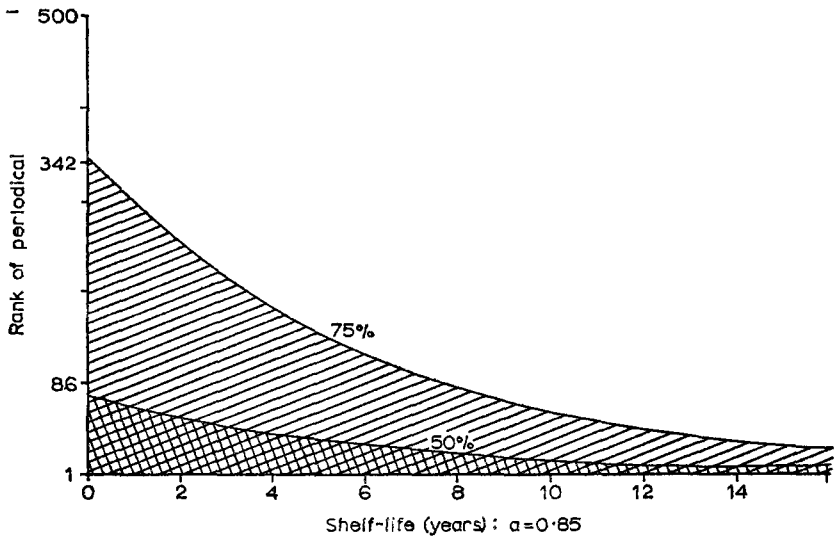

FIG. 6. The Contour Cut: Periodicals Discarded at Threshold Utility

into insignificance. Thus the literature of current scientific interest can be represented as the difference between an exponential growth and an equal exponential decay, but the growth line is well defined whereas the decay line is not. So the literature of *current interest* continues to grow but only as the difference between the curve of growth and the parallel but delayed curves of decay (figure 6).

The Optimum Specified Percentage of Utility for a Library of Specialist Periodicals

In this application Bradford's law is combined with the obsolescence law in the design of the most compact shelf-stock of periodicals which will provide any specified percentage of the utility obtainable from the acquisition and retention of the Bradford totality of N periodicals all with their complete book-runs.[10]

In the calculation that follows, it is assumed that all the relevant papers in the current issues of the complete set are of equal significance. This assumption is unlikely to be valid, for reasons mentioned earlier in this paper, but any error implicit in this assumption will yield a result which is on the conservative side.

If we now consider the utilities of the periodicals ranked as a Bradford set, we note that the productivity of the periodical ranked m (which lies outside the nuclear zone) is proportional to N/m. Thus the total utility of this periodical will be proportional to Nu/m. As all the periodicals age at the same rate, the more productive of two periodicals will always have the higher residual utility at any age t. If the library acquires the first m of the N ranked periodicals, then the periodical of rank m is at the minimum acceptable utility. The principle now applied is that the other periodicals are discarded when their residual utility declines to this minimum utility, i.e., equal residual utilities are discarded in the tails of each periodical.

For any other periodical of rank n, where $n \leq m$, we then note that it is discarded after t years where

$$Nua^t = Nu/m$$

which yields

$$a^t = n/m$$

so that the value of t can be read from the obsolescence graph.

The total utility lost by discarding the tails of the m periodicals will be $m\ Nu/m = uN$. The utility lost by not acquiring the periodicals of

rank $(n + 1)$ to N and discarding the periodical of rank m almost immediately will be

$$u(N \log N/s - N \log m/s) = uN \log N/m.$$

We can therefore write

$$uN + uN \log N/m = (1 - p)uN \log N/s$$

which expresses the fact that the utility discarded is equal to the fraction $(1 - p)$, where $p = P/100$, of the total utility of the complete Bradford set with all periodicals having full back runs. This equation yields

$$\log m/s = 1 + p \log N/s \tag{8}$$

from which m can be found for any specified value of p and known values of N and s.

Figure 7 shows the patterns of the $P\%$ library for $P = 50, 75$ when $N = 500$, $s = 2$, and $a = 0.85$. The nuclear periodicals are worth retaining without limit.

The "contour cut" just described is the most compact library, but it may not be the least costly because some of the periodicals of rank approaching m are discarded soon after acquisition. The simplest specified percentage of utility for a library is obtained by acquiring the requisite number of periodicals and discarding nothing. The appropriate number of journals in this case, m_1, is given by

$$N \log m_1/s = pN \log N/s$$

or

$$m_1 = N^p s^{1-p}. \tag{9}$$

Comparing (9) with (8) we see that $m_1 = m/e$ where e is the base of natural logarithms. So fewer periodicals are acquired but much more shelf space (and maintenance) is needed. The least costly solution is some compromise between these two solutions which depends on the relative costs of acquiring and of retaining and servicing the collection.

Although the contour cut solution may seem to be elaborate, it merely formalizes the way in which large collections of periodicals are dealt with practically in busy but efficient libraries where the users are allowed access to open shelves and where shelf space is limited. In such cases, the solution is determined by careful observation of usage rather than by measurement of obsolescence and application of Bradford's law. The volumes displaced from the open shelves are not discarded but are removed to closed-access stores nearby from which they can be re-

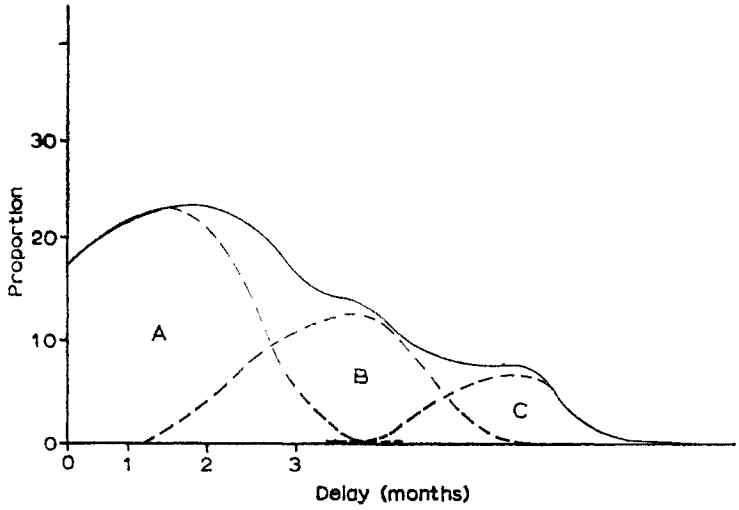

FIG. 7. Currency of SDI Services: A, National, B, Other, Same Language, C, Other, Different Language

trieved on request by members of the library staff. The problem here is therefore to insure that the volumes on open access are those most frequently sought by users. The principle applied is to remove to closed access those volumes whose average usage declines to some minimum frequency.

MEASURING THE EFFECTIVENESS OF RETROSPECTIVE SEARCHES AND CURRENT AWARENESS SERVICES

EFFECTIVENESS OF RETROSPECTIVE SEARCHES

Innumerable measures for quantifying the characteristics of information retrieval systems were proposed during the period of developing indexing languages and retrieval techniques during the 1950s and early 1960s. These measures were first applied to experimental manual systems with relatively small files and used with simple coordination of index terms which gave rise, for example, to Cleverdon's "inverse law of recall and precision."[11] Then Salton,[12] during his pioneering work on the mechanization of retrieval systems, introduced *normalized* recall and precision and the idea of ranking the items of the search output in terms of various measures of relevance. During this period, too, Swets[13] proposed theoretical models of information retrieval systems which were

again based essentially on the coordination of individual index terms drawn from a controlled list of terms.

Towards the end of that period the application of the computer to large data files became operational. And since that time, marked by the initiation of MEDLARS, the development of information systems has moved away from its obsession with the perfecting of index languages on sets of carefully prepared test-bed collections of well-indexed documents into an operational phase of development. Although indexing is still done mainly by technicians, the formulations of search questions have moved away from simple coordination of index terms into a more mechanical phase in which more sophisticated search languages, including syntactic components, such as the Boolean logical constants, have been developed. Step by step the computer has mechanized operations previously performed by human operators; first, the relatively easy matching of a group of terms representing the search question against groups of terms representing the documents; second, the elaboration of matching techniques, and now the replacement of the more intellectual processes of indexing by mechanized statistical techniques applied to titles and abstracts. The ultimate objective is the elimination of human operators, apart from the computer technicians, a task that will almost certainly be completed when input devices suitable for reading texts directly into the machine become more versatile, and indexing in terms of statistical analysis by machine of natural language texts becomes the norm.

In this phase of rapid technological development based on very large files and complex matching procedures, measures, empirical laws and theories derived from manual procedures and coordinate indexing have become increasingly inapplicable. The well-known average limits to effectiveness as measured by recall and precision seem to have become accepted as inescapably inherent in natural language descriptions. The only measure that remains directly applicable is *precision*, because that is the only measure that can be applied to the outputs of large mechanized operational systems. But precision by itself can be very misleading. Moreover there has been a shift of emphasis in operational systems from retrospective searching to current awareness services which are well suited to mechanization. As these powerful mechanized systems become established, new kinds of measure are needed; they will be more concerned with "user satisfaction" than with the efficiency of the searching mechanism itself.

The effectiveness of the retrospective searches of large mechanized

information systems is likely to be measured (unfortunately) in terms of their comprehensiveness. And, as a general test of the comprehensiveness of a bibliography, the application of the Bradford law, described earlier in this paper, is the only one yet available.

EFFECTIVENESS OF CURRENT AWARENESS SYSTEMS

Techniques of measuring user responses to current awareness (SDI) services are being developed by Leggate[14] and others. So far these techniques require the user to indicate which items he finds important, relevant but not important, and nonrelevant. Results of this kind can then be averaged and assessed using various conventional statistical techniques, although no single method has yet emerged which commands general acceptance. After all, the most critical test of operational systems run on a commercial basis is whether they can attract subscriptions from their users on a scale which makes them commercially viable.

One of the problems of concern to users of SDI services is that of currency. Users naturally expect the computerized SDI services they pay for to be up to date in their printouts. The user does not expect to see items which are a year or more old or which he knows have been superseded. The user of SDI services expects good currency rather than comprehensive coverage. Therefore, measures of currency are needed.

A practical difficulty in measuring currency is that the date of publication of a reference cited in an SDI printout, which is the nominal date of publication, may bear little relation to the date at which the reference was actually available in published form. So, in analyzing currency, the dates of the references should be measured, not from that quoted in the reference, but from the date of receipt of the source which, in an efficient library, is stamped on the source and corresponds to the day it was received. Random samples of the SDI output items can then be examined in relation to the date of availability and a frequency distribution of the delay in terms of months can be prepared. From such a statistical table two measures—the mean and the standard deviation—can be calculated, and both should be as small as possible.

Few data concerning currency have been published, but on the basis of the evidence available it seems that currency data take the form of truncated normal distributions as can most easily be shown by plotting the cumulative sums of delay times on arithmetic probability paper. If the SDI system offers coverage of worldwide sources it is interesting to separate the items which are yielded by the national sources of the country of the SDI service from those items which come from other

countries or from sources in other languages. In such cases one is liable to find two or even three overlapping normal distributions with increasing means and standard deviations (figure 7). Until the SDI system can accept direct inputs from the foreign sources, this aspect of currency is difficult to overcome; books and journals take appreciable time to be transported around the world, to be indexed or to be translated before abstracting and indexing.

One other irritant arises from the delayed publication of conference papers. These papers may bear the date of the conference but they may not be published in book or journal form for a year or more, and so they only then become available to the SDI service. Probably the solution to this difficulty arises from recognition of the different needs of retrospective searching and of current awareness. For comprehension the delayed conference papers need to be entered in the data base to be available for retrospective searches. But it is not necessary to compile the retrospective data base only by cumulating the periodic SDI collections which provide the current awareness services. The two services do not need to depend on precisely the same inputs.

Further Applications of Numerical Analysis

Goffman's general theory of information systems

The techniques of measurement applied to bibliography which have been described in the earlier sections of this paper have immediate practical application in the design of economic and efficient special libraries and bibliographical networks. But these measurements are also leading to quantitative descriptions of the characteristics of bibliographies which are gradually giving shape to what was regarded, only a few years ago, as an amorphous immeasurable confusion in which the seemingly limitless exponential growth of publication was the only quantified aspect. These quantitative descriptions are needed for the development of theories about scientific communication. If information science is to become the science it explicitly announces itself to be, it needs to develop a theory which lifts the subject off its present ground of ad hoc technological development and gives it autonomy, depth and more creative objectives.

A new area of analytical interest has been opened up by Goffman,[15] who has been exploring in detail the analogy between the dissemination of scientific knowledge and the spread of epidemics. He has analyzed the complete bibliographies of several emergent scientific descriptions

over periods of 100-200 years. He has shown that some of the ideas were "endemic" with minor outbreaks occurring from time to time. A man with some new knowledge is rather like the carrier of an infectious disease, but the disease is transmissible only to "susceptibles" and there must be a critical density of susceptibles for an epidemic to occur. The importance of this work lies in three directions: (1) it provides a new empirical analysis of the growth of science which should illuminate the recent philosophical descriptions of this process; (2) it offers a means of predicting whether an epidemic growth of an emergent subject is likely to occur or not; and (3) it establishes a basis for a much-needed general theory of communication which should give direction to the development of the worldwide information systems that are now being rather superficially discussed.

The theory Goffman is creating rests on two established theories—on epidemiological theory and on the Shannon information theory[16]—which are applied together in the analysis of the information phenomena recorded as comprehensive scientific bibliographies. The understanding of this theory requires competence in mathematics and statistics and cannot be summarized within the scope of this present article. It is mentioned here because it is already providing spin-offs of immediate practical interest. One of Goffman's needs is the compilation of comprehensive scientific bibliographies which in itself is an exercise of practical interest since it reminds us that the existing computerized data banks can offer little help with this project and that there is a need for a world repository of the tapes on which these hard-won bibliographies are now being recorded in FAMULUS format. Such tapes, suitably tagged for all foreseeable kinds of bibliographical analysis, are needed to increase the number of susceptibles that Goffman's own epidemiological theory of scientific growth requires for the epidemic growth of information science.

THE COMPACTION OF BIBLIOGRAPHIES

Although the computer has come to the aid of the bibliographer as a mechanism for compiling, editing, storing, and sorting the items that constitute the bibliography, it has done nothing to help in the task (which is not strictly that of the bibliographer) of sifting the grain from the chaff. The present ideal of the computerized data base is also that of the bibliographer—to compile *everything*. But if that means that the scientist in hot pursuit of some scientific objective is always going to be supplied with *everything* that is "relevant" whenever he asks for a retro-

spective search, he is going to be stopped in his tracks. The growth of science is all too often regarded as the steady accumulation of all knowledge, a view which is symbolized by the ever-growing data banks of the operational computerized systems. But. in fact, science is the continuous reappraisal of the old in terms of the new which discards the old almost as fast as it generates the new.

Until computerized data banks can reflect this more realistic view, their scientific users will become increasingly disillusioned with the lengthening printouts they receive. Work is therefore underway to develop methods of compiling, filtering and compacting scientific bibliographies so that current information can be presented in an immediately useful assimilable form. Information on tape ages at the same rate as the same information stored in archival collections of printed periodicals. No useful purpose is served by the continued revival of obsolete material, even when it is done by computer.

APPENDIX

The Relation Between Growth and Obsolescence

It is known that a geometric sequence is obtained from the citation dates obtained from a homogeneous literature whether the literature is growing or not. So, in general,

$$U_0 = R_0/(1 + a + a^2 + \ldots + a^{t-1} + \ldots) = R_0/(1 - a). \tag{1}$$

If the literature is growing exponentially at the rate g per annum, then after time t, R_t is given by

$$R_t = R_0 e^{gt}. \tag{2}$$

If the number of contributors (i.e., users) is also growing exponentially but at rate s per annum, then

$$U_t = U_0 e^{st}. \tag{3}$$

Differentiating (1), (2) and (3) with respect to t, substituting from (2) and (3) and simplifying, we have

$$\frac{da}{dt} = (s - g)(1 - a). \tag{4}$$

Clearly, if $s = g$, the value of a, as measured by citation counts, should remain constant. If $s \neq g$, then, on integrating (4), we have

$$(1 - a_t) = (1 - a_0)e^{(g-s)t} \tag{5}$$

where a_0 is the value of a when $t = 0$.

This result implies that while $g \neq s$ the value of a_t will change from year to year.

In the only empirical test of (5) yet published, Oliver[17] found that s and g *were* equal (within the limits of sampling error) for the literature of solid

state physics and that $a = 0.79$ at both the beginning and the end of a five-year period of rapid growth.

References

1. Bradford, Samuel C. *Documentation.* London, Crosby Lockwood, 1948.
2. Wilkinson, Elizabeth A. "The Ambiguity of Bradford's Law," *Journal of Documentation,* 28:122-30, June 1972.
3. Brookes, B. C. "Bradford's Law and the Bibliography of Science," *Nature,* 224:953-56, Dec. 6, 1969.
4. Goffman, William, and Morris, Thomas G. "Bradford's Law and Library Acquisitions," *Nature,* 226:922-23, June 6, 1970.
5. Brookes, B. C. "Photocopies v. Periodicals: Cost-effectiveness in the Special Library," *Journal of Documentation,* 26:22-29, March 1970.
6. ———. "The Design of Cost-effective Hierarchical Information Systems," *Information Storage and Retrieval,* 6:127-36, June 1970.
7. ———. "Obsolescence of Special Library Periodicals: Sampling Errors and Utility Contours," *Journal of the American Society for Information Science,* 21: 320-29, Sept. 1970.
8. Buckland, Michael K. "Are Obsolescence and Scattering Related?" *Journal of Documentation,* 28:242-45, Sept. 1972.
9. Naranan, S. "Power Law Relations in Science Bibliography: A Self-consistent Interpretation," *Journal of Documentation,* 27:83-97, June 1971.
10. Brookes, B. C. "Optimum P% Library of Scientific Periodicals," *Nature,* 232: 458-61, Aug. 13, 1971.
11. Cleverdon, Cyril W. "The Cranfield Tests on Index Language Devices," *Aslib Proceedings,* 19:173-94, June 1967.
12. Salton, Gerard. *Automatic Information Organization and Retrieval.* New York, McGraw-Hill, 1968.
13. Swets, John A. "Information Retrieval Systems," *Science,* 141:245-50, July 19, 1963.
14. Leggate, P. *Evaluation of Operational Current Awareness Services* (OSTI Report 5111). 1971. (Deposited in NLLST)
15. Goffman, William. "A General Theory of Communication." *In* Tefko Saracevic, ed. *Introduction to Information Science.* New York, R. R. Bowker, 1970.
16. Shannon, C. E. "The Mathematical Theory of Communication," *Bell Telephone System Journal,* 27:379-423, 623-58, 1948.
17. Oliver, Merrill R. "The Effect of Growth on the Obsolescence of Semiconductor Physics Literature," *Journal of Documentation,* 27:11-17, March 1971.