General Bibliometric Models

JOHN J. HUBERT

Introduction

OVER THE PAST fifty years, a sizable body of literature dealing with bibliometric models has developed. The early models were proposed because they were observed to fit graphically certain specific empirical frequency distributions. In many cases their functional forms were identical, the similarity only noted by other writers years later. In each case, depending on the subject field they applied to, there was a proliferation of papers which modified, extended, clarified, applied, and generalized the initial model.

Almost all bibliometric models relate, in a simple functional form, one variable with another variable. For example, in journal productivity studies, for a bibliography covering a certain span of years on a particular subject, a few journals contribute a large number of articles, other journals contribute fewer, and so on in a monotonic sequence ending with a large number of journals contributing one article each to the subject. The two variables are number of journals and number of articles. After arranging the journals in a decreasing order of productivity, a frequency-size distribution is obtained for the number of journals containing a fixed number of articles each. Conversely, a frequency-rank table can be constructed for the number of articles associated with a journal of fixed rank. These two approaches to observed patterns form the two modes of the data tabulations.

John J. Hubert is Associate Professor, Department of Mathematics and Statistics, University of Guelph, Ontario.

To illustrate explicitly the notions of the frequency-size approach, consider the following example. In table 1, f(n) denotes the number of journals contributing exactly n articles each to a particular subject field such that the total number of observed journals is $J = \Sigma f(n)$ and the total number of observed articles is $N = \Sigma nf(n)$. This tabulation relates the observations (the articles) with a class (a journal). The modeling problem is to find a mathematical equation relating f(n) with n. Associated problems are: What is the process which generates this relationship? What happens to the relationship if a larger sample of observations, N, is obtained? Does the relationship remain the same from year to year?

TABLE 1
A FREQUENCY-SIZE DISTRIBUTION OF THE NUMBER OF JOURNALS f(n) CONTRIBUTING n ARTICLES EACH

n	$f(n)$	$nf(n)$
1	102	102
2	25	50
3	13	39
4	2	8
5	7	35
6	1	6
7	3	21
8	3	24
9	1	9
10	2	20
13	2	26
15	1	15
18	1	18
22	1	22
Sum	J=164	N=395

Source: S.C. Bradford. "Sources of Information on Specific Subjects," *Engineering* 137(1934):85-86.

In the last twenty-five years, it has been observed that such tabulations occur for other pairs of variables from a wide variety of natural and social phenomena. Table 2 provides some examples of such combinations of observation versus class relationship.

To understand the frequency-rank approach, consider the example given in table 1. Near the bottom of the table there is one journal contributing the most (twenty-two) articles. This journal is assigned the rank 1. The next most productive journal is assigned rank 2 because it

TABLE 2
EXAMPLES OF OBSERVATION-CLASS RELATIONSHIP

Observation	*Class*
Number of articles	journals
Number of citations	persons
Number of insects	species
Length of word	words
Number of papers	authors
Number of occurrences	initial digits
Checked-out frequency	books
Number of occurrences	nouns
Length of sentence	sentences
Number of phonemes	words
Income level	persons

contributed eighteen articles. This is continued, resulting in the frequency-rank distribution given in table 3, where g(r) is the number of articles contributed by the journal of rank r. Notice that there are two journals contributing thirteen papers each, and each is assigned rank 5, the "maximal-rank" assignment method which is used in the case of ties. (If we assign the rank 4 to each of these journals, then we are using a "minimum-rank" method; there are also the random-rank and average-rank methods.) The frequency-rank tabulation reverses the order of the frequency-size tabulation, and gives priority to the most productive journals. The frequency-size approach gives emphasis to the journals of least productivity. There are other relationships between the two approaches. Advantages and disadvantages of the frequency-rank approach are discussed by Hubert and others.[1]

For the examples given in table 2, the literature contains many models, and some are erroneously referred to as "laws" as if they predicted occurrences without error. From an analysis of these models, it becomes apparent that some are for the frequency-size approach and some are for the frequency-rank approach. The modeling problems have different purposes, because from the data in table 1 the model can be used to predict the number of journals contributing a fixed number of articles, and from the data in table 3 the model can be used to predict the number of articles contributed by a journal of a given rank. An explanation of the list of all the different models which can be found to be applicable to bibliometric phenomena, including the actual equation, the variables each relates to, the approach to obtain the equation, and how they interrelate, would be extremely lengthy and beyond the

TABLE 3
A FREQUENCY-RANK DISTRIBUTION OF THE NUMBER OF ARTICLES g(r) CONTRIBUTED BY A JOURNAL OF RANK r

r	$g(r)$
1	22
2	18
3	15
5	13
7	10
8	9
11	8
14	7
15	6
22	5
24	4
37	3
62	2
164	1

present scope and purpose of this article. However, each article in the appendix to this paper contains a model which would be included in this list because each adequately fits and models some form of tabulation. One word of caution is necessary: some of the models have been declared as new and general, while others are self-declared and are neither new nor general. There are survey articles on many of these models, and some of these articles provide the mathematical equations, historical developments, interrelationships, and examples of data sets where the models have been useful.[2]

There are three models which are claimed to be general because they possess two important properties: first, they include earlier models as special cases; and second, they are applicable to a large class of bibliometric variables. These are the models of Price, Bookstein and Brookes. Bookstein especially has claimed that the major bibliometric models—Bradford, Lotka and Zipf—are in fact "a single law that seems capable of describing phenomena in a vast variety of subject areas."[3] The three models of Price, Bookstein and Brookes are discussed in the following sections, with special attention to their derivations and to their appropriateness as general models that can account for some of the individual models mentioned above.

Analysis of the Price Model

The Price model[4] is also known as the cumulative advantage distribution (CAD) and can be defined as follows: if f(n) is the fraction of contributors having n articles each, then $f(n) = (m+1)\,B(n, m+2)$, for $n = 1, 2, \ldots$, with the parameter $m > 0$, and $B(\bullet, \bullet)$ is the Beta function. The Beta function is a name for a fundamental integral* involving two parameters, and there is no simple verbal expression for this function.[5] The CAD was proposed as a frequency-size type model because it yields the relative frequency or proportion of authors each of whom has produced a fixed number of articles on a specific area over a fixed period of time. Over a finite range of observational values of n, a distribution of authors is obtained, and the model can be fitted so as to follow closely the observed pattern. When the fit is statistically adequate it can be used, for example, to predict the percentage of authors who have contributed more than n papers each, and if n is large, this provides an estimate of the set of so-called prolific authors on a subject area. Other important uses such as in citation analysis have been illustrated by Price.

This model has as a rough approximation that f(n) is proportional to n^{-a}, where $a > 0$. This implies that as n increases, f(n) decreases, which suggests that there are many authors having one paper each, and so on in a decreasing fashion, with very few authors contributing many papers. There is only one parameter in the model, and its value depends on a particular data set. Price himself considers his model to be quite general: "It provides a sound conceptual basis for such empirical laws as the Lotka Distribution for Scientific Productivity, the Bradford Law for Journal Use, the Pareto Law of Income Distribution, and the Zipf Law for Literary Word Frequencies. It is therefore an underlying probability mechanism of widespread application and versatility throughout the social sciences."[6]

How does one obtain such a model? The early attempts before 1950 by Yule, Pareto, Zipf, and Bradford were based on plotting the data with f(n) versus n, for example, then finding a mathematical equation which would adequately represent the pattern observed in the particular discipline (Yule in biology, Pareto in economics, Zipf in linguistics, and

---

*The Beta function is also known as Euler's first integral and is defined as:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{(a-1)!\,(b-1)!}{(a+b-1)!}, \qquad a > 1,\ b > 1,$$

where $n! = n(n-1)\cdot\ldots\cdot 3\cdot 2\cdot 1$, if n is an integer. Also, B(a,b) is approximately proportional to a^{-b} under certain conditions.

Bradford in journal productivity). In 1955, Simon derived the basic form of the Price model, and proved it was a consequence of two assumptions.[7] If a collection of N articles is found on a specific subject area, and if f(n) represents the number of journals containing n articles each, then in this bibliometric framework, the two assumptions are: (1) the probability that the next article found in a journal which already has contributed n articles is proportional to nf(n), the total number of occurrences of all articles from those journals which already have n articles each on the subject area under study; and (2) there is a constant probability that the next article found is from a new journal. These assumptions form the basis of what is known as the stochastic birth or growth process.

Although the derivation by Simon is very rigorous and the statistical theory used is very advanced, it does result in the same model equation that Price proposed twenty years later. Simon also established the model's generality by showing that it contains: (1) the models of Yule and Willis in biology, (2) the models of Zipf and Mandelbrot and others in linguistics,[8] (3) the models of Zipf in population growth, and (4) the models of Pareto and Champernowne in income distributions.

The two assumptions of Simon are plausible, relatively simple and satisfy many social processes; however, there is one drawback: they are not unique, because other mechanisms can be shown to lead to the same model equation. One of two other starting points is due to Simon himself, and the other is, in fact, the Price starting point. These two starting points will be considered separately.

Simon's second mechanism, in journal productivity terminology, is as follows: Suppose we have a collection of N articles dispersed among J journals such that f(n) represents the number of journals contributing n articles each. Furthermore, suppose articles are added to the collection according to the two assumptions of the former growth process, and articles are dropped from the collection in such a way that the sample size N remains constant. Simon then proves that the same model equation involving the Beta function can be derived if we assume that if an article from a particular journal is dropped, then all articles from that journal are dropped, and the probability that the next journal dropped be one contributing exactly n articles is proportional to f(n). This added assumption will account for articles leaving or entering the collection, i.e., the processes of emigration and immigration. It also can be used to mimic changes in distributions due to different time periods but constant sample sizes.

The Price starting point which generates this model equation involving the Beta function is a modification of the classical "Polya

urn" scheme. Suppose the contents of an urn containing two types of colored balls depend upon what was selected in previous draws. If a ball of the first color is drawn (called a "success"), two or more balls of that same color are replaced so that on the next draw there is an increased chance of obtaining a ball of that color. The modification occurs when a ball of the second color is drawn, in which case a single replacement of that color is made so that in the next draw the chance of drawing this second color is not increased. The net effect is that success increases the chance of further success, whereas failure has no effect in changing the chance of success or failure.

The success-breeds-success concept has some empirical evidence to support it, e.g., in the sociological theory of publishing characteristics, in citation analysis, and in usage patterns from retrieval systems in libraries, as well as in biological and epidemic processes. Therefore, what Price has accomplished is to begin at a different starting point (the urn scheme) and end at the same final model equation as Simon did, who started with the birth process assumptions.

In summary, the Price model equation involving the Beta function has the following properties: (1) it is a frequency-size model; (2) it has the limiting form that f(n) is proportional to n^{-a}, for some constant $a > 0$; (3) it approximates several models in the literature; (4) it is the same as the model proposed by Simon; and (5) it can be derived from three different starting points, two due to Simon and one due to Price. Therefore, although Price's theory underlying the model is sound and new, the model equation and its ability to described bibliometric phenomena has been known since 1955. However, as a model equation it is general because it satisfies our definition involving the two conditions: it must model different variables, and it must contain or approximate earlier models. It is interesting to note that the theory surrounding this model equation is not entirely complete: "The surface has only been scratched and doubtless the application of this theory will raise more empirical testing and rigorous statistical mathematics in expression."[9]

Analysis of the Bookstein Model

In 1977 Bookstein proposed to find an expression for the expected number of authors, f(n), in a discipline producing *n* articles over a defined period of time, subject to sociological factors influencing productivity and other constraints.[10] The factors used were society's need for research and the use of "rewards and threats" for continued productivity. There were two constraints; the first was that Lotka's model be a

special case. (Lotka's model is also known as the inverse-square law, and essentially states that f(n) is proportional to $1/n^2$, for n = 2, 3,... .) The second constraint is that if a publication distribution is observed over t time periods (e.g., t = 10 years), then the function f should satisfy the relation $f(tn) = f(t) \times f(n)$. Bookstein calls this the "symmetry property" or the "invariance property."[11] Bookstein claims that the only realistic function satisfying these conditions and empirical data is f(n) proportional to $1/n^a$ where a is a positive number and estimable from the data. (It is true that for this model equation we have Lotka's law when a = 2, and furthermore, the symmetry property is satisfied since $f(tn) = 1/(tn)^a = (1/t^a)(1/n^a) = f(t)f(n)$.) It is also claimed that the model is the only one which is unchanged whether the population of authors under study remains the same, increases or decreases over time.[12] This claim has not been convincingly demonstrated.

There are four important observations which can be made about this model:

1. The model equation is a special case of the model equation involving the Beta function advocated by both Simon and Price. In fact, Bookstein recognizes this: "Simon's model and mine...are not identical, they converge at large n."[13]
2. The model equation is not the only possible equation satisfying his two constraints.
3. The path to the model is different from the other paths discussed earlier. In 1924 Yule used the empirical data fitting technique; in 1955 Simon used stochastic birth process assumptions; in 1976 Price used the urn scheme mechanism; and in 1977 Bookstein used symmetry and other conditions to establish the model.
4. The model is not original. The form of the Bookstein model equation appears in earlier papers, as demonstrated in Fairthorne and Hubert,[14] where we see that the very early models of Pareto, Zipf and Stevens, and later Naranan[15] are exactly this model for the frequency-size tabulation. Hubert has proposed this same model equation for the frequency-rank tabulation.[16]

The implication of the first observation is that the Bookstein model is a special case of the model involving the Beta function. Therefore, in this sense, the Bookstein model is less general. Also, since the model involving the Beta function fits many observable variables, because it is so adjustable to a variety of shapes, and since the form n^{-a} is not as adjustable, then, in this sense, the Bookstein model is less general. We will return to the property of generality in a later section.

Analysis of the Brookes Model

In 1977 Brookes claimed to have proposed a model which is "...an empirical law of social behaviour which pervades all social activities" and for which "Bradford's law can be regarded as a particular example." Also, Brookes believes in "...the wide generality of the Bradford law."[17] This section considers the models of both Bradford and Brookes since they are apparently related.

In 1934 Bradford stated his famous model after examining how 395 articles on lubrication were dispersed among 164 different journals.[18] The actual data are given in table 4, where G(r) is the total number of articles in the first r most productive journals. The Bradford model is $G(r) = a + b \log(r)$, where $r = 1, 2,...$ and a and b are parameters depending on the subject area. When the cumulative totals of articles are plotted against the logarithm of r an almost straight-line relationship results. This approach gives priority to the most productive journals. When tables 3 and 4 are compared, it is clear that the variable r is the same. This is the reason the Bradford model is called a ranking type of model. Brookes argues that this model can be used in other social contexts whenever sources of an activity are ranked in order of decreasing activity. This approach of ranking is very important to Brookes: "*Ranking by frequency* is a technique widely used and understood....Ranking is more primitive than measuring. We learn to rank before we learn to speak or count. It is *because* ranking is a primitive action which permeates all social activities that it is time it were taken more seriously."[19] It is probably true that papers on bibliometric modeling refer more to the Bradford model than to any other model. We will not digress further on the Bradford model, but consider the Brookes model.

The structural form of the model proposed by Brookes is much more complicated than the Bradford model: if g(r) is the number of references in the rth most productive journal, then

$$g(r) = \frac{k}{j!} \sum_{j=r}^{\infty} \frac{m^j}{j!}$$

where $r = 1, 2,...$, $m > 0$ is a parameter, k is a quantity depending on m, and $r! = r(r-1)(r-2)\times...3\times2\times1$. Unfortunately, this equation has no simple verbal or mathematical expression, but it does possess several properties which clarify its form:

1. The variable r acts as a rank because it is equivalent to the maximum-rank assignment scheme mentioned earlier.

TABLE 4
THE BRADFORD-TYPE TABULATION OF THE ACCUMULATED NUMBER OF REFERENCES G(r) CONTAINED IN THE FIRST r MOST PRODUCTIVE JOURNALS

Accumulated No. of Journals r	*Accumulated No. of References* $G(r)$
1	22
2	40
3	55
5	81
7	101
8	110
11	134
14	155
15	161
22	196
24	204
37	243
62	293
164	395

2. The mathematical properties are proper since the infinite series converges, $g(r) \to 0$ as $r \to \infty$ and $g(1) \geq g(2) \geq \ldots$, i.e., monotonicity.
3. The model relates the number of references, g(r), with the rank r, whereas the Bradford model relates the cumulative number of references, $G(r) = \Sigma_{s=1}^{r} g(s)$, with the rank r; that is, the Brookes model is a frequency function and the Bradford is a distribution function.
4. When m is large and when we consider cumulative totals, the Brookes model does conform to the Bradford model, i.e., $\Sigma_{s=1}^{r} g(s) \simeq a + b \log (r)$.
5. The model gives priority to the most productive journals because the journals with only a few articles are in the tail of the frequency function.
6. The model is based on the well-known Poisson discrete random variable which also possesses a countable infinte number of values.
7. The model is adjustable to a variety of shapes.
8. The model is entirely new, and its exact structure is not like any other model.

Brookes calls his model "the mixed Poisson model" because the derivation depends on a mix of Poisson random variables. In general terms, the mix occurs as follows: for the sum $X_1 + X_2 + \ldots + X_n$ we assume

not only that the Xs are independent Poisson random variables, but also that n, the number of variables, is a Poisson random variable. This is the concept of "random sum of random variables" instead of a fixed sum of random variables. More specifically, the underlying assumptions of the Brookes model can be reduced to the following: (1) the number of articles produced by a journal per unit time is a Poisson random variable with mean rate of, e.g., θ; and (2) the total number of journals, each producing at mean rate θ, is inversely proportional to θ. The second assumption is consistent with the observation that as the rate of production increases, the number of journals decreases, or the most productive journals (lowest rank numbers) produce the greatest numbers of articles. The derivation is therefore based on realistic assumptions.

Another interesting consequence of Brookes's model is his modifications of the Bradford model. Earlier, Brookes proposed a hybrid form for the Bradford model to account for the nonlinearity at the beginning of observed distributions.[20] He suggested the modified Bradford model:

$$G(r) = \begin{cases} \alpha r^{\beta}, & r = 1, 2, \ldots, c, \\ a + b \log r, & r = c+1, c+2, \ldots, n. \end{cases}$$

Notice that for $r = 1, 2, \ldots, c$ the function is a curve, and for large values the function is a straight line function of $\log r$. To conform to Brookes's new model and other observed distributions, he now suggests two hybrids, called Type I and Type II, which he claims take the form:

$$G(r) = \begin{cases} \log_b [(a + \sum_{j=0}^{r} \alpha^{c-j})/a], & r = 1, 2, \ldots, c \\ \log_b [(a + r)/a], & r = c+1, c+2, \ldots, n, \end{cases}$$

where $b = (a+n)/a$ and $\alpha < 1$ for Type I and $\alpha > 1$ for Type II. Graphically, these functions appear in figure 1, where hybrid Type I is convex initially and hybrid Type II is concave (with respect to the r-axis) initially. The hybrids are consequences of his model and illustrate its ability to adjust to anomalies.

In summary, the Brookes model is included in this article because of its properties and its declared generality. To quote Brookes: "The main advantage of the model is that it shows how the log law, and therefore how the hybrid forms of the Bradford law, can be derived in a realistic and natural way from orthodox frequency statistics"; and "in its present form it is the simplest possible stochastic model of the Bradford law, but it can easily be modified, for example, to embrace

problems of growth and obsolescence—the classical 'birth and death' process of stochastic theory."[21]

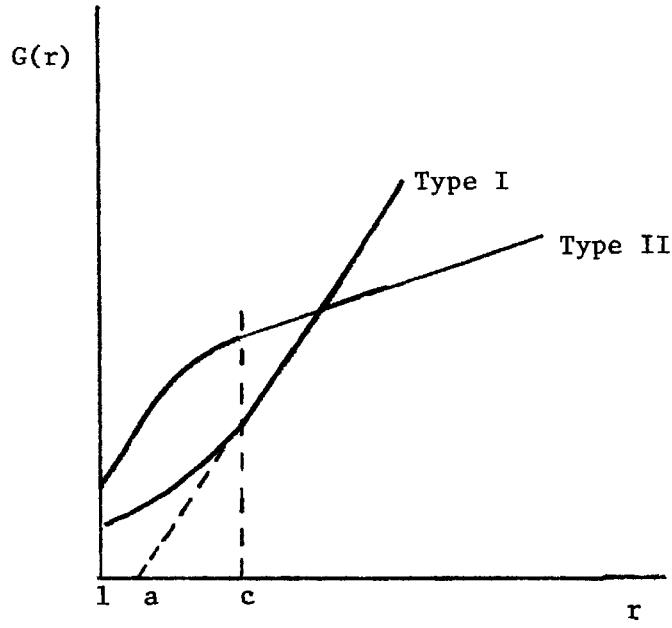

Fig. 1. The Brookes hybrid types of Bradford's model
Source: Bertram C. Brookes. "Theory of the Bradford Law." *Journal of Documentation* 33(Sept. 1977):193.

The Validity of the Generalizations

Let us now return to the question of whether the models of Price, Bookstein and Brookes are valid general models. It should be stressed that the structural form of the Brookes model is new, but the Price and Bookstein models are not new. We have shown that the Price model was first proposed by Simon in 1955 and that the Bookstein model has been proposed by many others.[22] However, we have explained how the assumptions underlying the models are original and indeed helpful in the understanding of the processes which could generate the models.

With respect to their generality, it has been demonstrated that all three models possess the two properties of the original criterion, that is,

they include earlier models as special cases, and they are applicable to a larger class of bibliometric variables. However, these general models are limited in that they consider only the effect of one variable upon another. Nature and life are not so simple. In fact, in bibliometrics, recent articles have attempted to model one response variable as a function of two or more variables. Also, on one source (journal, author, etc.) more than one response variable has been measured. These two approaches will change our definition of generality because such multivariate models will necessarily include the univariate models. It is a simplistic viewpoint of reality to believe one variable in a social interactive process can be adequately predicted solely by one other variable. A univariate model does not become more general by merely including more parameters.

Examples of models of greater statistical sophistication can be found: Bayesian models in interactive and retrieval systems,[23] methods for evaluating articles,[24] stochastic literature growth models,[25] modeling duration of book usage,[26] measures of literature concentration using the Whitworth model in frequency-rank distributions,[27] modeling relationships between title length and number of coauthors,[28] properties of modeling,[29] and prediction models using time-series methods.[30]

This latest research differs from earlier work in bibliometrics in that it uses models that are nonlinear and that consider the effect of several variables, i.e., they are multivariate. These models require the estimation of at least two parameters, whereas the simpler univariate models required only one. The maximum likelihood method, the minimum chi-square method, and the ordinary linear least-squares method have been used. However, estimation for nonlinear functions requires care. If a model is linear and of the form $Y = \alpha + \beta X + \epsilon$ (where the random variable ϵ must have structure if confidence limits are to be established), we speak of an additive model for the variable Y depending on the variable X. If $Y = \alpha X^{\beta}\epsilon$, then this is an example of a multiplicative model. Taking logarithms on both sides, we have $\log Y = \log \alpha + \beta \log X + \log \epsilon$, which is of the form $Y = \alpha' + \beta X' + \epsilon'$. We have "linearized" the model where $\epsilon' = \log \epsilon$ has a lognormal structure. For the nonlinear model $Y = \alpha X^{\beta} + \epsilon$, taking logarithms yields $\log Y = \log (\alpha X^{\beta} + \epsilon)$, which does not collapse into a linear form. This simple fact is often overlooked, and the estimation of parameters for such models requires nonlinear estimation theory.[31]

The use of multivariate models also requires greater care. If Y is found to be functionally dependent on p variables $X_1, X_2, \ldots, X_p$, such as $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$, then we have a multiple regression

model. If the response on a single subject is a set of variables $Y_1, ..., Y_n$, which may be correlated and are functionally dependent on a set of variables $X_1, ..., X_p$, then we have a multivariate regression model. The latter situation can utilize techniques such as cluster, factor and multivariate time-series analyses. Although recent articles in retrieval systems are using time-series methodology, the simpler models listed earlier in this article are not multivariate, and it should be possible to exploit multivariate methods to achieve clarity and more generality.

Summary

The frequency-size and frequency-rank approaches, the two basic approaches in a class of bibliometric models, have been explained. The twenty-eight known models have been cited, and the three models due to Price, Bookstein and Brookes have been analyzed by considering their internal properties, interrelationships and generality. Because they have a sound but different statistical foundation, they possess validity; however, except for possibly Price's model, it is clear that the models are not used in everyday prediction problems in library and information science. Also, it has been shown that the Price and Bookstein models are not new. The three models are of limited generality because they are univariate and simple. Examples of more sophisticated models have been cited, and remarks have been made to suggest how greater generality can be achieved by using multivariate methods.[32]

References

1. Hubert, John J. "Analysis of Data by a Rank-Frequency Model." Ph.D. diss., State University of New York at Buffalo, 1974; Brookes, Bertram C., and Griffiths, Jose M. "Frequency-Rank Distributions." *Journal of the ASIS* 29(Jan. 1978):5-13; Hubert, John J. "Bibliometric Models for Journal Productivity." *Social Indicators Research* 4(Oct. 1977):441-73; and _____. "A Relationship Between Two Forms of Bradford's Law." *Journal of the ASIS* 29(Jan. 1978):159-61.

2. Simon, Herbert A. "On a Class of Skew Distribution Functions." *Biometrika* 42(Dec. 1955):425-40; Brookes and Griffiths, "Frequency-Rank Distributions"; Price, Derek de Solla. *Little Science, Big Science*. New York: Columbia University Press, 1963; Fairthorne, Robert A. "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction." *Journal of Documentation* 25(Dec. 1969):319-43; Brookes, Bertram C. "Theory of the Bradford Law." *Journal of Documentation* 33(Sept. 1977):180-209; Bookstein, Abraham. "The Bibliometric Distributions." *Library Quarterly* 46(Oct. 1976):416-23; and Hubert, "Bibliometric Models."

3. Bookstein, Abraham. "Explanations of the Bibliometric Laws." *Collection Management* 3(Summer/Fall 1979):151-62.

4. Price, Derek de Solla. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the ASIS* 27(Sept.-Oct. 1976):292-306.

5. Ibid.
6. Ibid., pp. 292-93:
7. Simon, "On a Class of Skew Distribution Functions."
8. *See* Hubert, John J. "Linguistic Indicators." *Social Indicators Research* 8(June 1980):223-55.
9. Price, "A General Theory," p. 304.
10. Bookstein, Abraham. "Patterns of Scientific Productivity and Social Change: A Discussion of Lotka's Law and Bibliometric Symmetry." *Journal of the ASIS* 28(July 1977):206-10.
11. Ibid., p. 208; and _____, "Explanations of the Bibliometric Laws," p. 159.
12. _____, "Patterns of Scientific Productivity," pp. 206-10.
13. _____, "Bibliometric Distributions," p. 422.
14. Fairthorne, "Empirical Hyperbolic Distributions"; and Hubert, "Bibliometric Models."
15. Naranan, S. "Power Law Relations in Science Bibliography—A Self-Consistent Interpretation." *Journal of Documentation* 27(June 1971):83-97.
16. Hubert, "Analysis of Data."
17. Brookes, "Theory of the Bradford Law," p. 180.
18. Bradford, S.C. "Sources of Information on Specific Subjects." *Engineering* 137(26 Jan. 1934):85-86.
19. Brookes, "Theory of the Bradford Law," p. 203.
20. Brookes, Bertram C. "The Derivation and Application of the Bradford-Zipf Distribution." *Journal of Documentation* 24(Dec. 1968):247-65.
21. _____, "Theory of the Bradford Law," pp. 185, 202.
22. Fairthorne, "Empirical Hyperbolic Distributions"; and Hubert, "Bibliometric Models."
23. Tague, Jean M. "A Bayesian Approach to Interactive Retrieval." *Information Storage and Retrieval* 9(March 1973):12-42; Bookstein, Abraham, and Cooper, William. "A General Mathematical Model for Information Retrieval Systems." *Library Quarterly* 46(April 1976):153-67; and Inhaber, H. "Canadian Scientific Journals: Part II, Interaction." *Journal of the ASIS* 26(Sept.-Oct. 1975):290-93.
24. Virgo, Julie A. "A Statistical Procedure for Evaluating the Importance of Scientific Papers." *Library Quarterly* 47(Oct. 1977):415-30.
25. Braun, Tibor, et al. "Literature Growth and Decay: An Activation Analysis Résumé." *Analytical Chemistry* 49(July 1977):682-88.
26. Cooper, Michael D., and Wolthausen, John. "Misplacement of Books on Library Shelves: A Mathematical Model." *Library Quarterly* 47(Jan. 1977):43-57.
27. Pratt, Allan D. "A Measure of Class Concentration in Bibliometrics." *Journal of the ASIS* 28(Sept. 1977):285-92; and Carpenter, Mark P. "Similarity of Pratt's Measure of Class Concentration to the Gini Index." *Journal of the ASIS* 30(March 1979):108-10.
28. Kuch, T.D.C. "Relation of Title Length to Number of Authors in Journal Articles." *Journal of the ASIS* 29(July 1978):200-02.
29. Rouse, William B. "Tutorial: Mathematical Modeling of Library Systems." *Journal of the ASIS* 30(March 1979):181-92.
30. Kang, Jong H., and Rouse, William B. "Approaches to Forecasting Demands for Library Network Services." *Journal of the ASIS* 31(July 1980):256-63.
31. *See, for example*, Wold, Herman. "Nonlinear Estimation by Iterative Least Squares Procedures." In *Research Papers in Statistics*, edited by Florence N. David, pp. 411-44. New York: Wiley & Sons, 1966.
32. Research for this paper was partially supported by NSERC Grant No. A9229.

Appendix

Articles Containing Models of Bibliometric Phenomena

Benford, Frank. "The Law of Anomalous Numbers." *Proceedings of the American Philosophical Society* 78(1938):551-72.

Bookstein, Abraham. "Patterns of Scientific Productivity and Social Change: A Discussion of Lotka's Law and Bibliometric Symmetry." *Journal of the ASIS* 28(July 1977):206-10.

Bradford, S.C. "Sources of Information on Specific Subjects." *Engineering* 137 (26 Jan. 1934):85-86.

Brookes, Bertram C. "The Derivation and Application of the Bradford-Zipf Distribution." *Journal of Documentation* 24(1968):247-65.

_____, and Griffiths, J.M. "Frequency-Rank Distributions." *Journal of the ASIS* 29(1978):5-13.

Cole, P.F. "A New Look at Reference Scattering." *Journal of Documentation* 18(June 1962):58-64.

Goffman, William, and Newill, Vaun A. "Generalization of Epidemic Theory; An Application to the Transmission of Ideas." *Nature* 204(17 Oct. 1964):225-28.

Good, I.J. "Distribution of Word Frequencies." *Nature* 179(16 March 1957):595.

_____. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika* 40(Dec. 1953):237-64.

Harris, Bernard. "Determining Bounds on Integrals with Application to Cataloging Problems." *Annals of Mathematical Statistics* 30(June 1959):521-48.

_____. "Statistical Inference in the Classical Occupancy Problem Unbiased Estimation of the Number of Classes." *Journal of the ASIS* 63(Sept. 1968):837-47.

Herdan, Gustav. *Type-Token Mathematics: A Textbook of Mathematical Linguistics.* The Hague: Mouton, 1960, pp. 182-85.

Hubert, John J. "Analysis of Data by a Rank-Frequency Model." Ph.D. diss., Dept. of Statistics, SUNY-Buffalo, 1974.

Kendall, Maurice G. "Natural Law in the Social Sciences." *Journal of the Royal Statistical Society, Series B* 124(1961):1-16.

Leimkuhler, Ferdinand. "The Bradford Distribution." *Journal of Documentation* 23(Sept. 1967):197-207.

Lotka, A.J. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* 16(1926):317-23.

Naranan, S. "Power Law Relations in Science Bibliography—A Self-Consistent Interpretation." *Journal of Documentation* 27(June 1971):83-97.

Pareto, Vilfredo. *Cours d'Économie Politique.* Lausanne: F. Rouge & Cie., 1896. *See esp.* vol. 2, sec. 3.

Plackett, R.L. "The Truncated Poisson Distribution." *Biometrics* 9(Dec. 1953):485-88.

Price, Derek de Solla. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the ASIS* 27(Sept.-Oct. 1976):292-306.

Rao, I.K. Ravichandra. "The Distribution of Scientific Productivity and Social Change." *Journal of the ASIS* 31(March 1980):111-22.

Resnikoff, H.L., and Dolby, J.L. *Access: A Study of Information Storage and Retrieval with Emphasis on Library Information Systems* (Final Report HEW Proj. 8-0548, 1972).

Simon, Herbert A. "On a Class of Skew Distribution Functions." *Biometrika* 42(Dec. 1955):425-40.

Stevens, S.S. "On the Psychophysical Law." *Psychology Review* 64(1957):153-81.

Vickery, B.C. "Bradford's Law of Scattering." *Journal of Documentation* 4(Dec. 1948): 198-203.

Willis, John C. *Age and Area; A Study in Geographical Distribution and Origin of Species*. Cambridge, Eng.: University Press, 1922.

Yule, G. Udny. "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. John C. Willis, F.R.S." *Philosophical Transactions of the Royal Society, Series B* 213(1924):21-87.

Zipf, George K. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley, 1949.

This Page Intentionally Left Blank