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

ON BIBLIOMETRIC MODELING

Ferdinand F. LEIMKUHLER

Purdue University, West Lafayette, Indiana

Abstract

A common feature in bibliometric studies is the use of mathematical models to analyze fundamental problems arising from the operation of information systems. As bibliometrics develops, more explicit attention needs to be given to the modeling process as a unifying activity within the field, a vital link to other fields of study, and an avenue to future growth. In this paper the author draws on his experience with bibliometric modeling to demonstrate its practical and theoretical significance, and cites some recent computational developments that are revolutionizing modeling methodology.

1. MATHEMATICAL MODELING

The purpose of mathematical modeling is to analyze fundamental problems associated with the operation of real systems. Modeling provides norms for the design of such systems and the evaluation of system outcomes. It also provides important methodological links to the modeling expertise developed in other application areas. Within a specific area, modeling provides the framework for developing and defining a common body of knowledge that is characteristic of a particular scientific discipline.

As a practical way of solving operational problems and designing better systems, modeling can be viewed as one of four steps in the cyclic development process shown in Figure 1. The steps are problem evaluation, model analysis, system design, and operation control. In large organizations, these functions might be separated and highly specialized. In smaller ones the clarification of these distinct functions may be less clear and they might be done by the same person. In any case, they need to be done in close interaction with one another.

The model is related to operations indirectly through problems and systems. It is a way of representing problems in solution (or system) form and, at the same time, representing systems in generic (or problem) form. Thus, the model can be used to help system designers find meaningful options, and at the same time, help problem evaluators set goals and choose an optimal strategy for controlling operations. While problem evaluations and system design may occur without recourse to formal modeling, it is likely that both activities depend on implicit norms. Part of the modeler's task is to make norms explicit.

Formal modeling provides a valuable connection to the analytic expertise in other application areas. It is a communication channel for finding valuable insights and analogies in formulating problems that have common characteristics in different settings and promotes the development of a common methodology for modeling system problems. How a problem is modeled frequently depends more on the expertise of the modeler than the intrinsic nature of the problem being modeled.

In addition to its practical importance as a problem-solving and system-design tool and as a way of exploiting the expertise in cognate fields, modeling plays a very significant role in developing and defining new fields of study. In this regard, models are paradigms that incorporate laws, theories, rules, applications, and methods that define a particular coherent tradition of scientific research.

F.F. Leimkuhler

As described by T.S. Kuhn [1], paradigms play a crucial role the development of science. "In its normal state," Kuhn [1, p. 166] says, a "scientific community is an immensely efficient instrument for solving the problems or puzzles that its paradigms define".

2. LIBRARY STORAGE MODEL

A fundamental library problem is how to best organize a collection so as to maximize the collection's usefulness and minimize its cost. Stringent space constraints and unusual book dimensions may make it necessary and reasonable to deviate from uniform subject classification schemes. Serials are usually separated from monographs. Items that receive very heavy use or very little use may be shelved differently. Several examples are given below of attempts to model these kinds of problems, develop better shelving and classification schemes, and evaluate their relative advantages. Also, these examples are intended to show how the modeling of one aspecti of library operations can be related to parallel studies in other areas and thereby make it possible to develop richer paradigms for the study of an inherently complex activity.

Shelving by size in libraries dates back at least to Melvil Dewey's [2] recommendation of limited shelving to accompany his decimal classification plan. The literature on shelving reflects the practicality of the subject and often is lacking in experimental data or reasons for the claims made. Even Fremont Rider's [3] exhaustive study of shelving lacked hard evidience for his conclusions. J.G. Cox and the author [4] proposed the model shown in Figure 2, *where a cumulative height distribution function F(H) represents the fraction of* lineal shelving required for books of relative height or less $0 \leq H \leq 1$. If only one shelf size is used, it must be of relative height $H = 1$ and the required area is 1. If two shelf heights are used, with $H = 1$ and $H = h < 1$, then the total presented area is A(h), where $A(h) = hF(h) + 1\text{-}F(h)$, and the saving in space is 1-A(h), as shown in figure 2. The optimal value of h can be found by the dynamic programming methods. Experiments with actual book height date from the Purdue Libraries indicated the results shown in Table 1.

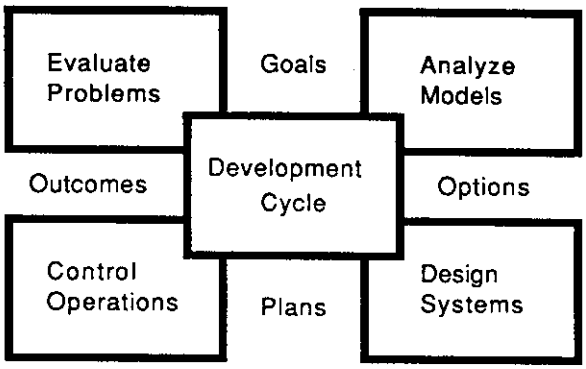

Figure 1 : The Problem-Model-System-Operation Cycle of development

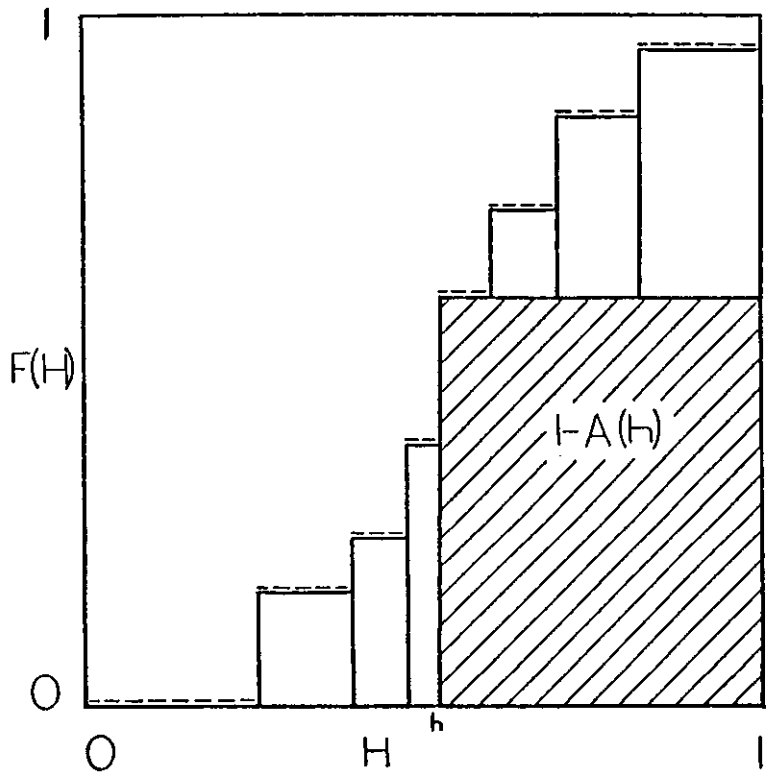

Figure 2 : Cumulative height distribution F(h) and area 1-A(h) saved if shelf height of relative size h and 1 are used.

Table 1 : Percent Capacity Increase

Number of Size Classes	Shelve by Height	Shelve by Width
1	0 %	27 %
2	38 %	94 %
3	47 %	111 %
4	51 %	116 %
5	53 %	121 %
10	58 %	128 %

3. LIBRARY CLASSIFICATION MODEL

While storage by size may be useful in large depository facilities, most libraries are more concerned with segmenting collections according to their potential use. An idealized model for representing such a system was proposed by the author [5] and is represented in Figure 3 for a collection that has been ranked according to its "usefulness" for a user search. In Figure 3, the curve F represents the cumulative probability that the search is successful if the collection is ordered according to the probability of success for each item. The diagonal line F(1) on the other hand, represents the cumulative probability of success if all items are equally successful, i.e., random search. The area above this line, which equals 0.5, is the expected search length. The difference in area above F(1) and F is the gain in expected search between time and ordered and unordered collection.

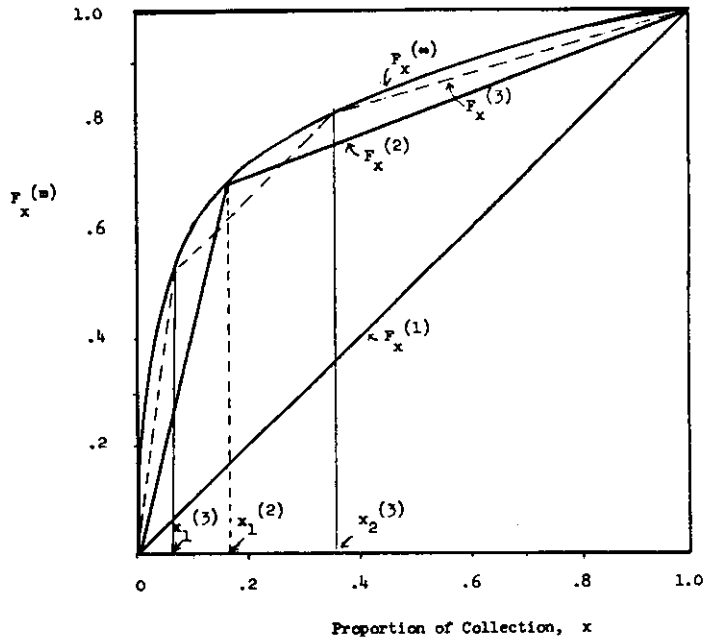

Figure 3 : Cumulative probability of success F(m) for a search with m zones. Items in each zone have the same probability which is less than that of the previous zone.

Frequently subject classification schemes divide collections into two zones : those that are more useful and those less useful. If one searches randomly in these two groups, with the more useful group first, the effectiveness of this strategy is represented by the triangular line F(2) with $x_1(2)$ percent of the collection being in the more useful zone. A three-way division of the collection is represented by the broken line F(3) with $x_1(3)$ percent of the collection in the first zone of most useful items and, $1-x_2(3) - x_2(3)$ percent in the third zone of least useful items.

By formulating the curve F and using suitable optimization techniques, one can find optimal values for dividing the collection. For example, if the curve has the form of a Bradford-type distribution with a parameter value of 269, the most efficient two-zone search is to have about 17.5 % of the most productive items in the first zone. This would result in a mean search effort over 24 % of the collection as compared to an effort over 17.5 % if the collection were perfectly ordered in decreasing order of potential usefulness. An optimal three-zone search would examine, first, the most productive 7.5 % of the collection, second, the next most productive 28.5 %, and then the remainder of the collection. The expected search effort would require examination of 20 % of the collection and end in the second zone.

The above models examples demonstrate how shelving and classification models can be used to help evaluate storage and retrieval problems. They measure the relative benefit of subdividing a collection according to some attribute of size or use and show how to do this optimally. The models can be extended to other kinds of information system problems. For example, when variable length computer records are recorded in fields of fixed length, the shelving model has been modified to find optimal field lengths for a file.

In addition to practical applications, classification models can help establish a basic algorithmic framework for further bibliometric research. Progress in applied science is made by identifying fundamental paradigms and then exploring all of their implications. The bibliometric laws of Lotka, Zipf, and Bradford constitute such a paradigm set, and there is a meaningful connection between these laws and the shelving and classification models. The appearance of a particular paper in a particular journal is the result of a classification decision. Mandelbrot [9] has shown how the efficient coding (classification) of messages can lead to a Zipf distribution of word frequencies.

4. STOCHASTIC MODELS

Numerical experimentation with the classification models depends on a formulation of the distribution of probable usefulness. It was for this purpose that the author [6] derived a linear-logarithmic formulation of Bradford's law.

Classification models add another dimension to the large literature on bibliometric laws and the persistent problems of estimation and validation that are associated with skew distributions. Recently, Y.S. Chen and the author [7] showed that by assigning an index to each observation of a ranked set of bibliometric data, the essential equivalence of Lotka's law, Zipf's law, and Bradford's law could be demonstrated without recourse to goodness-of-fit methods. However, fundamental estimation problems remain because of the absence of the properties normally assumed in statistical analysis. Furthermore, the traditional bibliometric laws provide only a statistic representation, at a particular point in time, of the output of an on-going dynamic process, and there is a need to develop stochastic models of information processes.

Markov [8], Mandelbrot [9], and Simon [10] have proposed models that focus on the dynamics of information processes. Markov's model is based on the assumption of probabilistic regularity in the recurrence of words in a string of words. Mandelbrot's model assumes that words are selected according to Shannon's theory on the optimal coding of messages in communication systems. Simon [11, p. 62] rejected these mechanistic approaches, preferring a more humanistic approach based on tendencies toward "imitation" and "association" in

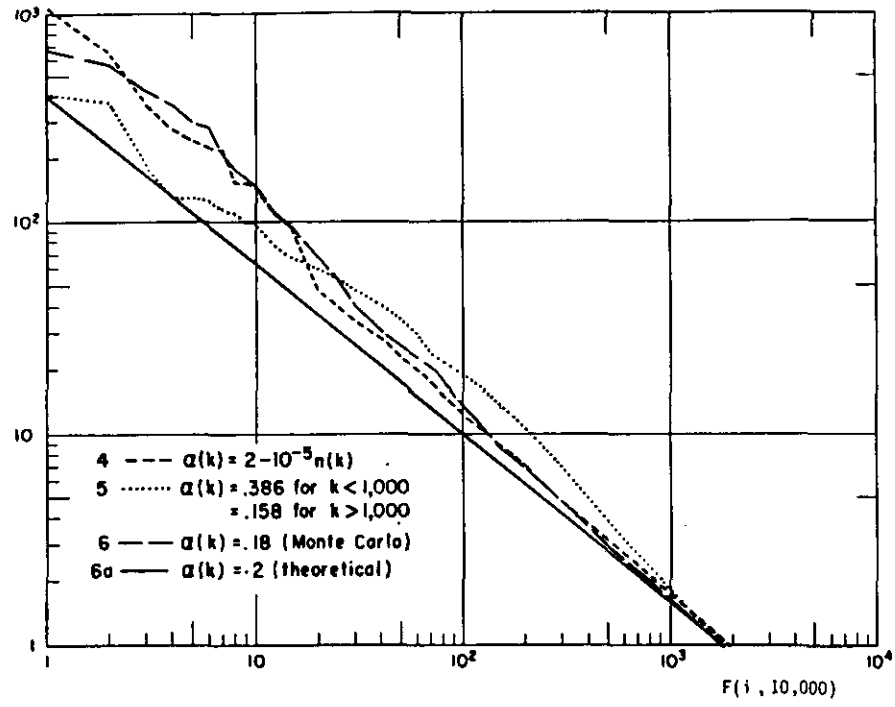

Figure 4 : Monte Carlo estimates [11] of the frequency distribution of a Simon-Yule process with alpha close to 0.2.

the choice of models of expression. Simon's model makes on two relatively simple assumptions about the appearance of words in a string : (a) there is a probability alpha that the next word is a new word that has not appeared before; and (b) if not new, the old words recur in proportion to their prior frequency of appearance. This process can be easily programmed on a computer to simulate skew distributions like those shown in Figure 4. Simon argues that the process generates approximate Yule distributions that follow Zipf's law.

Y.S. Chen [12] made an extensive comparison of the Markov, Mandelbrot, and Simon models and concluded that the Simon-Yule models appear to be the most promising for informational studies. In addition to generating marginal distributions that approximate the bibliometric laws, the Simon-Yule model exhibits two other interesting properties. First, the interval between the successive appearances of the same word are exponentially distributed. Second, the total number of "types" of words is related to total number of all words or "tokens" according to a "type-token" ratio observed in many empirical studies of text.

The development of stochastic bibliometric models adds considerable complexity to the problems of verifying, validating, and experimenting with such models. Simon [11, p. 134] discusses these problems at length, and proposes the following steps.

1. Begin with raw data, not theories.
2. Draw simple generalizations from striking features in data.
3. Find limiting conditions by manipulating the variables.
4. Devise simple mechanisms to explain steps 2 and 3.
5. Propose explanatory theories that go beyond step 4 and make experiments.

Simon notes that statistical theory is not very helpful in this process.

4. COMPUTATIONAL EXPERIMENTATION

In a recent collection of essays on "The Craft of Probabilistic Modeling," M.F. Neuts [13] notes that two key issues confront the further development of science along the path suggested by Simon. The mathematical community has been slow to accept "experimentation" as an important aspect of applied mathematics.

Consequently, much of the work in computational experimentation is being done by persons who are motivated by the practical benefits of using Monte Carlo simulations to justify the design of complex systems, without sufficient regard for the authenticity of the results. A second problem is the fact that this methodology requires a radical shift in viewpoint from traditional concerns with "average behavior" to the study of systems under "extreme conditions". Neuts [13, p. 221] says :

> In order to obtain the necessary information for a thorough understanding of such models, one often needs to apply a laborious and hybrid methodology. Next to such theorems as can be proved, one needs to compute, not one, but several probability distributions and this for a variety of parameter values. *The numerical results* require cogent interpretation--which is often difficult--and, in some cases, confirmation by experimentation with realistic data. Major fields of technological and medical endeavor are now viewing stochastic modeling as highly significant and useful. The problems they bring are difficult and non traditional, and I have no doubt that algorithmic analysis will play a major role in their solution. As so often in science, ultimate rewards will go to those who dare to seize the day and help shape the future.

In another paper, Neuts [14] elaborates on the nature and need for computer experimentation as the only plausible way to study real systems under extreme conditions.

Such studies make it necessary to go beyond classical analytical methods to a new kind of algorithmic and experimental mathematics. Future progress will require the development of a new set of rules for conducting experiments, new collections of exemplary studies, and new standards for validating results. The use of computer experimentation is growing rapidly in many areas of application, and particularly in the newer areas of artificial intelligence, expert systems, and knowledge engineering. These delopments plus the early identification by Simon of the need for such methods in information studies makes them seem particularly relevant to further development of paradigms in the field of bibliometrics and their application to information system design and problem solving.

REFERENCES

[1] Kuhn, T.S., The Structure of Scientific Revolutions, (University of Chicago Press, Chicago, IL, 1970).
[2] Dewey, M., Library Shelving : Definitions and Principles, Library Notes 12 (1887) p. 105-106.
[3] Rider, F., Compact Book Storage, (The Nadham Press, New York, NY, 1959).
[4] Leimkuhler, F.F., Cox, J.G., Compact Book Storage in Libraries, Operations Research 12 (1964) p. 419-427.
[5] Leimkuhler, F.F., A Literature Search and File Organization Model, Amer. Doc. 19 (1968) p. 131-136.
[6] Leimkuhler, F.F., The Bradford Distribution, J. of Documentation 23 (1967) p. 197-207.
[7] Chen, Y.S., Leimkuhler, F.F., A Relationship between Lotka's law, Bradford's law, and Zipf's law, J. of Amer. Soc. for Info. Sciences 37 (1986) p. 307-314.
[8] Markov, A.A., An Example of Statistical Investigation of the Text of Eugen Onegin Illustrating the Connection of Trials in a Chain. Bull. de L'Acad. Imperiale des Science de St. Petersburg 7 (1913) p. 153.
[9] Mandelbrot, B., An Information Theory of the Statistical Structure of Language, Proc. of Symp. on Applications of Communication Theory (Butterworths, London, 1953).
[10] Simon, H.A., On a Class of Skew Distribution Functions. Biometrika 52 (1955) p. 425-440.
[11] Ijiri, Y., Simon, H.A., Skew Distributions and the Sizes of Business Firms, (North Holland, Amsterdam, 1977).
[12] Chen, Y.S., Statistical Models of Text, (Ph.D. Thesis, Purdue University, 1985).
[13] Neuts, M.F., An Algorithmic Probabilist's Apology in : Gani, J. (ed), The Craft of Probabilistic Modelling (Springer-Verlag, New York, 1986).
[14] Neuts, M.F., Computer Experimentation in Applied Probability, (Working Paper 86-030, Systems and Industrial Engineering Dept., University of Arizona, 1986).