Information Processing & Management Vol. 19, No. 3, pp. 173–186, 1983
Printed in Great Britain.

0306-4573/83/030173-14$03.00/0
Pergamon Press Ltd.

MATHEMATICAL MODELS OF INFORMATION SYSTEM USE†

ROBERT M. HAYES and HAROLD BORKO
Graduate School of Library and Information Science University of California, Los Angeles, CA 90024, U.S.A.

(*Received for publication* 19 *November* 1982)

Abstract—This report presents the results from a study of mathematical models relating to the usage of information systems. For each of four models, the papers developed during the study provide three types of analyses: reviews of the literature relevant to the model, analytical studies, and tests of the models with data drawn from specific operational situations.

(1) The Cobb–Douglas model:

$$x_0 = a x_1^{\,b} x_2^{(1-b)}.$$

This classic production model, normally interpreted as applying to the relationship between production, labor, and capital, is applied to a number of information related contexts. These include specifically the performance of libraries, both public and academic, and the use of information resources by the nation's industry. The results confirm not only the utility of the Cobb–Douglas model in evaluation of the use of information resources, but demonstrate the extent to which those resources currently are being used at significantly less than optimum levels.

(2) Mixture of Poissons:

$$x_0 = \sum_{i=0}^{n} i \sum_{j=0}^{p} n_j \, e^{-m_j} (m_j)^i / i!$$

where x_o is the usage and $(n_j, m_j), j = 0$ to p, are the $p + 1$ components of the distribution. This model of heterogeneity is applied to the usage of library materials and of thesaurus terms. In each case, both the applicability and the analytical value of the model are demonstrated.

(3) Inverse effects of distance:

$$x = a \, e^{-md} \text{ if } c(d) = rd$$

$$x = a d^{-m} \text{ if } c(d) = r \log(d).$$

These two models reflect different inverse effects of distance, the choice depending upon the cost of transportation. If the cost, $c(d)$, is linear, the usage is inverse exponential; if logarithmic, the usage is inverse power. The literature that discusses the relationship between usage of facilities and the distance from them is reviewed. The models are tested with data from the usage of the Los Angeles Public Library, both Central Library and branches, based on a survey of 3662 users.

(4) Weighted entropy:

$$S(x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} r(x_i) p(x_i) \log(p(x_i)).$$

This generalization of the "entropy measure of information" is designed to accommodate the effects of "relevancy", as measured by $r(x)$, upon the performance of information retrieval systems. The relevant literature is reviewed and the application to retrieval systems is considered.

†This research was supported by the National Science Foundation, Division of Information Science and Technology, Under Grant No. IST 79-18497.

1. INTRODUCTION

The project presented in this report and associated papers (see Section 8, Publications from the Mathematical Models Project) was designed to test and validate a set of mathematical models relating to various aspects of the usage of information systems. One objective was to provide a better understanding of the variables affecting the usage of information systems. Beyond that, a second objective was to provide, where possible, design equations on which to base both economic and technical decisions about the structure and operation of information systems.

1.1 *Background*

There have been a number of qualitative surveys and statistical studies of information system usage and there have been many mathematical models proposed to describe that usage. Perhaps the best known example is Zipf's law which, either in itself or in variant forms like Bradford's law or Lotka's law, has been used to describe the distribution of frequency of use among the information resources of a system.

Susan Crawford reviewed "user" studies which focus on the user as such[1]. She pointed out that the methods for such studies have developed well and have "produced both effective case studies and field studies". She goes on to say, "These studies utilized well designed survey instruments; carefully selected, stratified, random sampling; and appropriate techniques of statistical analysis". But she concludes "... (they) have restricted our ability to generalize and to develop theory". Another growing area of research is "bibliometrics", part of which is a quantification of usage as it relates to the records stored in the information system[2]. In particular, Bradford's early recognition of the concentration of publication certainly reflects this kind of focus. Studies in this area have been highly quantitative, as is natural, since they deal with easily measurable phenomena, and they have a solid basis in theory.

This project lies at the intersection of these two kinds of study, the area in which the concern is with the interaction between the user's behavior and the operation of the information system. Specifically, four contexts of user/system interaction were considered, for each of which a mathematical model was used to represent hypotheses underlying that interaction. The purpose was therefore to test the hypotheses, if possible to validate the models, and to explore the consequences with the aim of developing design equations. More fundamentally, the objective was to identify parameters relevant to a theory which will explain and predict the relationships between information resources and their utilization.

1.2 *The models*

(a) *Allocation of resources.* The first context was the economic relationship between usage and decisions made about allocation of resources. The model proposed to represent the relationship was a standard econometric model, the Cobb–Douglas model:

$$x_0 = a x_1^{b} x_2^{(1-b)}$$

where x_0 is the measure of production, x_1 is the expenditure for operating expenses (especially for labor), and x_2 is the expenditure for capital.

The hypothesis was that this model would apply to information system contexts, with the variable x_2, normally interpreted as capital, being now interpreted as investment in information resources. There was already some evidence to support that hypothesis, based on examination of the usage of public libraries (with production measured by circulation, labor by public service staff, and capital by the book collection plus the costs of technical processing). The consistency of results was high enough to warrant further test, on other groups of public libraries, other types of libraries, and other means for providing information services. For each test, it would be necessary to develop measures of production, operating costs, and capital investment that would be appropriate.

The model leads naturally to design equations for the optimum allocation of resources between labor and information, as the capital investment in information systems.

(b) *Distribution of usage.* The second context was the operational relationship between usage and file items. The model proposed to represent the relationship was the mixture of Poisson distributions:

$$x_0 = \sum_{i=0}^{n} i \sum_{j=0}^{p} n_j \, e^{-m_j} (m_j)^i / i!$$

where x_0 is the usage and (n_j, m_j), $j = 0$ to p, are the $p + 1$ components of the distribution.

It is well recognized that the usage of an information resource is not uniformly distributed, and Zipf's law has frequently been called on to represent the frequency distribution. However, there are several problems that arise when one tries to use Zipf's law (or its variants). First, it rarely fits actual distributions closely enough for acceptance of it as the underlying law with any degree of statistical confidence. Second, the differences quite typically occur at exactly those values of most significance (i.e. the very small or very large frequencies), the ones where management decision must be made. Third, while there is an underlying stochastic process that may lead one to consider Zipf's law (basically, that usage is governed by "contagion", or the assumption that usage increases likelihood of future usage) the law is expressed as a steady-state situation; it thus describes what the status of things is, not how it got that way. Fourth, there are many cases in which the underlying assumption of a contagion process is not applicable. As a result of these considerations, it was decided to model the distribution of usage of library resources by the mixture of Poisson distributions.

The model leads naturally to design equations, especially to determine the structure of the file system in terms of levels of accessibility. The special significance of the mixture of Poisson distributions is that it permits the design equations to recognize heterogeneity in the usage.

(c) *Effects of distance.* The third context was the relationship between usage and the distance between the user and the information system. Two models were proposed to reflect different inverse effects of distance, the choice between them depending upon the cost of transportation:

$$x = a \, e^{-md} \text{ if } c(d) = rd$$

$$x = ad^{-m} \text{ if } c(d) = r \log(d).$$

Theoretical studies of the use of public facilities, as a function of distance, have suggested that the functional relationship is dependent upon the cost of transportation in these ways. If the cost $c(d)$ is linear, then the usage is inverse exponential; if logarithmic, the usage is inverse power. In either case, "distance" can be interpreted very broadly to include psychological as well as physical distance.

These models have direct consequences in design equations to determine the optimum spatial distribution of service facilities and point of access to an information system.

(d) *Weighted entropy.* The fourth context is the relationship between usage and the amount of information provided. The model proposed as the measure of the amount of information was the "weighted entropy" measure, considered as a generalization of both the entropy measure, of communication theory, and the relevance measure that traditionally has been used for the evaluation of retrieval systems:

$$S(x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} r(x_i) p(x_i) \log(p(x_i))$$

where $(x_1, x_2, \ldots, x_n)$ are the signals (or document records), $p(x_i)$ is the *a priori* probability of signal x_i, and $r(x_i)$ is the relevance of that signal.

It has been recognized that "uncertainty" is a crucial variable in the usage and performance effectiveness of an information system. A specific example is provided by

"inter-indexer consistency", with evident effects upon precision and recall. The classical means for dealing with phenomena like uncertainty is to apply the "entropy measure for information", in which the uncertainty is represented by "noise" in a communication channel, and the information capacity of the channel is a function of the signal to noise ratio. Coding theory provides means for achieving a desired reliability in communication, given the characteristics of the environment of noise.

The difficulty with the classical entropy measure is that it reduces the concept of information to the purely statistical properties of signals. This is not satisfactory, either from a psychological viewpoint or in trying to apply it to information retrieval contexts. The problem is that it fails to reflect the issue of "relevance", a concept crucial to retrieval, however, it may be measured; it fails to recognize the significance of the signals received to the user or recipient of them. Communication theory and the entropy measure treat all signals as equally relevant; the weighted entropy model accommodates the fact that signals differ in their relevance.

1.3 *Overview of this paper*

In the following section of this paper, we will summarize the significant results from the examination of these models. In each case, we will make reference to the related publications produced during the project, including some that are still "working papers", being readied for publication. Section 8 of this paper provides a complete listing of the publications to which reference is made.

2. APPLICATION OF THE COBB-DOUGLAS MODEL TO THE USE OF INFORMATION BY U.S. INDUSTRY

In this section, we discuss the studies related to the application of the Cobb–Douglas model to the relationship between information services and national productivity.

2.1 *Sources of data*

The data used in this portion of the study were obtained from the Office of Telecommunication, U.S. Department of Commerce. They are the same data on which Marc Porat based his study of the "information economy" of the United States[3]. Additional data were obtained from Rubin-Taylor[4].

2.2 *Information resources in the national economy*

The first area of study was focussed on the interaction between usage of information services and the allocation of resources in the national economy in general. As a first step, the relevant literature was reviewed, including publications of Machlup, Marschak, Porat and others, in order to establish a working definition of the "information economy". As the most recent and best documented, the definition used by Porat was adopted and used to analyze the relationship between information, treated as a capital investment, and productivity. The basis for the analysis is the Cobb–Douglas model, taken in the log–linear form:

$$\log x_0 = \log a + b \log x_1 + (1 - b) \log x_2$$

where x_0 is a measure of production, x_1 a measure of labor, and x_2 a measure capital investment.

Application of the Cobb–Douglas model to U.S. industry was then used to interpret the results which Porat describes as "information overhead"—the effects of increasing allocations of resources to purchases of information. It shows that the results from the model are totally consistent with those from Porat. The real issue turns out not to be the apparent overhead, but the use of a production model as the basis for optimum allocation of resources. Full reports of this portion of the study appear in the papers by Hayes (1980) and by Borko (1981A).

2.3 *"Added value" as a function of information expenditures*

An analysis was undertaken to evaluate the extent to which investment in information resources can be demonstrated to have direct value in greater industrial profitability. Specifically, the study examined the relationship between "added value" for the manufacturing industries of the country and the investment they make in purchased information services.

Results from this part of the study are presented in Hayes (& Erickson) (1982A). In it, anecdotal evidence is discussed as the rationale for treating information as a component of productivity. Then the Cobb–Douglas model is considered as means for statistical verification of the effects of information purchases by industry. Theoretical conditions for optimum use of purchased resources, including information resources, are presented. Data describing the expenditures by each industry, for labor, capital, information purchases, and other purchases are then analyzed. The paper concludes with a discussion of the heterogeneous nature of the data and summarizes the results from analysis of a range of alternative models, all of which have the same qualitative results.

The results obtained can be summarized succinctly: The Cobb–Douglas model is applied to "value added" as a function of labor, capital, purchases of information services, and purchases of other input. A regression analysis for that model is applied to 50 manufacturing industries, using data from 1967, and to 51 industries, using data from 1972. The theory states that, if industry is operating in an optimal manner, the marginal return from the external purchases (whether for information services or for other input), as expressed by the coefficient in the regression, should be zero. If the coefficient is positive, industry is not using enough of the external resource; if negative, industry is using too much of it.

The following are the regression equations for 1967 and 1972 data (in which V is the added value, L is the labor force, K is capital resources, I is the purchases of information services, and X is the purchases of other inputs to production):

$$\log V = \log A + a \log L + b \log K + c \log I + d \log X, \quad \text{correlation} = \mathrm{R}$$

Year	Log A	a	b	c	d	R^2
1967	1.564	0.307	0.352	0.292	0.022	0.985
1972	1.504	0.256	0.415	0.320	−0.014	0.981

The near zero value of the coefficients for log X(purchases of other input) suggest that those resources are being used optimally. The large positive coefficients for log I (purchases of information services) suggest that they are being used far less than would be of optimal return to added value and therefore to profit.

In conclusion, the results show that there is a demonstrable relationship between increased added value (and therefore profitability) and investment in information resources, and that manufacturing industry is using far less then the optimal amount of information resources.

3. APPLICATIONS OF THE COBB–DOUGLAS MODEL TO LIBRARIES

In this section, we discuss the application of the Cobb–Douglas model to evaluation of the productivity of libraries, both public and academic.

3.1 *Sources of data*

(a) *Data bases related to public library studies.* Statistical data were obtained for public libraries in each of several states (California, Georgia, Maryland, Montana, New Jersey, New York, Ohio, Wisconsin), using the publications from the state library agencies for each of those states. Similar data were obtained from the U.S. Department of Education (LIBGIS data tapes) for the country as a whole.

(b) *Data bases related to ARL studies.* Data were taken from tapes compiled from statistics about the ARL libraries for the years 1968–1979, covering collection size, library staff, and similar characteristics of the library and the institution. These data were augmented with data about faculty and numbers of publications by them (as derived from the Source Indexes of the Social Science Citation Indexes for the years 1971–1981).

3.2 *Allocation decisions in libraries*

The literature relevant to formalized procedures for allocation of resources was reviewed. The papers by Palmer (1980) and by Borko (1981) summarize the results in terms of the following issues:

- Application of the Clapp–Jordan formula for collection adequacy (and modifications of it to meet perceived local needs) to allocation of resources in development of collections.
- Descriptive models based on regression analyses of variables describing library operations and services.
- Formulas for allocation of book budgets among academic disciplines, based on equity, importance, usage, and cost vs benefit.
- Formulas for allocation among different categories of materials (especially, serials vs monographs).
- Formulas for allocation among different categories of staff (technical processing vs public service).

3.3 *Application of Cobb–Douglas to public libraries*

The initial application of Cobb–Douglas to libraries was to public libraries in California and then to those in a few other states. Those results were reported in two papers, Hayes (1979) and Hayes (1980A).

For these analyses, labor was taken as the "Reader Service Staff" (calculated as the total staff less the "technical services staff", the latter being treated as part of the capital investment in the collection). The capital investment is represented by the size of the collection (it being the primary capital resource of the library, with other investments, such as the buildings to house the collection and the labor costs in selection and in cataloging, taken as proportional to the size of collection). The production of a public library is measured by circulation.

The data used for one analysis were all taken from the LIBGIS file, statistics collected by the Department of Education for the libraries of the country. The results of the regression on those data were as follows:

	Number	$\log(a)$	$1-b$	R
Budget < 1$ million	1260	0.294	0.842	0.69

where $\log(a)$ and $(1-b)$ are the parameters in eqn (1) and R is the correlation of the two variables in that equation.

A second set of analyses used the same equation, but used data reported for each of several states individually, the data being derived from the annual reports of each state library or other reporting agency. The results of these several regression analyses were as follows:

State	Number	$\log(a)$	$1-b$	R
California	120(of 173)	0.709	0.654	0.70
Illinois	454(of 567)	0.633	0.676	0.79
Ohio	230(of 251)	0.617	0.691	0.78
Missouri	112(of 121)	0.670	0.631	0.64
Montana	53(of 91)	3.680	0.641	0.73
New York(Counties)	60(of 62)	0.582	0.718	0.87

In each state (except New York, where county totals were used), the sample of libraries was limited to those with budgets of less than $1 million per year in order to eliminate those libraries or systems with major reference functions which are not necessarily well measured by circulation. In Montana, data were not available for several libraries.

The very similar values of $1 - b$ and R provide clear confirmation of both the general applicability of the Cobb–Douglas model to public libraries and to its applicability to specific groups of libraries, such as those in states.

3.4 *Criteria for optimum allocation*

Hayes (1979A) considers design equations for optimum allocation of resources. To do so, the Cobb–Douglas model, when applied to public libraries, must be used in two different forms, representing the two quite different contexts of a central collection and a distributed, branch oriented collection:

● For a Central Library, minimize the total budget,

$$T_C = c_1x_1 + c_2x_2$$

subject to satisfying the circulation demand, with the use of resources represented by the Cobb–Douglas model in the form:

$$x_0 = a(x_1)^b(x_2)^{1-b}$$

The resulting criteria for optimum allocation of resources then are

$$c_1x_1 = T_C(b)$$

$$c_2x_2 = T_C(1 - b)$$

● The optimum conditions for branch libraries must recognize the effect of distance upon the demand for services. To reduce distance, one increases the number of branches and thus the staffing requirements (assuming a branch must have at least a minimal number of staff, represented by m in the following):

Minimize

$$T_B = c_1x_1 + c_2x_2$$

Subject to

$$x_0 = a(B/B_0)(mB)^b(x_2)^{1-b}$$

The resulting criteria for optimum then are

$$c_1x_1 = T_B(1 + b)/2$$

$$c_2x_2 = T_B(1 - b)/2.$$

These conditions were applied to data from the Los Angeles Public Library and Los Angeles County Library branch systems. The value for the parameter $(1 - b)$ was taken as 0.64, representing the general average of the results from regressions on public libraries (as shown in the prior part of this section).

LAPL Branches	Theoretical Optimum	Estimated Actual
Reader Services Staff	$8,915,000 (68%)	$9,327,000
Collection Budget	4,196,000 (32%)	3,784,000
	13,111,000 (100%)	13,111,000
LA County System	Theoretical Optimum	Estimated Actual
Reader Services Staff	$13,166,000 (68%)	$13,112,000
Collection Budget	6,196,000 (32%)	6,250,000
	19,362,000 (100%)	19,362,000

In each case, the difference between optimum and actual is less than 4% of the total budget.

3.5 *Applicability of Cobb–Douglas to ARL libraries.*

The paper by Hayes (and Pollock & Nordhaus) (1982) considers application of the Cobb-Douglas model to research libraries. Two sets of analyses are presented: "library productivity" and "institutional productivity".

For the first analysis, labor is taken as the "Reader Service Staff" (calculated as the total staff less the "technical services staff", the latter being treated as part of the capital investment in the collection). The capital investment is represented by the size of the collection (it being the primary capital resource of the library, with other investments, such as the buildings to house the collection and the labor costs in selection and in cataloging, taken as proportional to the size of collection).

The difficulty lies in measuring the production of an academic research library. Whereas there is an easily obtainable, relatively reliable measure of production of public libraries, the situation for academic research libraries is much more complicated. While those libraries generally report circulation statistics, the evidence is clear (and will be discussed in more detail when we consider the analyses of the Distribution Model) that circulation is by no means an adequate measure of the total use of them. In particular, and of greatest importance, the research use of a major academic research library is different both in form and amount from circulation use.

In order to apply the Cobb–Douglas model, therefore, it is necessary to have surrogate measures for their production, and especially of production in support of the research functions of the university they serve. Erickson's paper in this part of the study thus focussed on "Estimating the production function for research libraries". The result was the identification of three surrogate measures of the research productivity of the institution and thus of the library in support of that research: (1) number of faculty; (2) number of Ph.Ds produced; and (3) number of publications—all three with respect to each institution.

To apply the Cobb–Douglas model, data were needed for the usual variables related to the library (collection size, staff, acquisitions, and budget), but data were also needed for additional variables related to the institution and covering the three measures of research productivity. Initially, it was hoped that the HEGIS file from the Department of Education would serve as a source for most of these variables, but that turned out to be an unreliable and otherwise unsatisfactory data base.

Therefore, data were acquired from the Reports of the ARL for the past ten year period; these provided the data on the variables related to the libraries. A variety of other sources were used to obtain data about the several ARL universities. Among them was the Social Science Citation Index, the source files of which were searched (online, using the ORBIT system of SDC) to obtain number of publications attributed to persons at each of the ARL universities for each of the past ten years.

Most of the data for that first analysis were taken from ARL Statistics. (The nature of that source of data is discussed, with emphasis on some of the inconsistencies in the data.) The data were then subjected to three regression analyses for the ARL libraries. The results for 1975/76, as a representative year among the eight to ten years covered by the analysis, are as follows:

Labor = Services Staff	$\log(a)$	$1-b$	R
x_0 = Ph.Ds produced	−0.76	0.85	0.676
x_0 = Number of Faculty	0.62	0.46	0.341
x_0 = Interlibrary loans	0.91	0.92	0.537

where $\log(a)$ and $(1-b)$ are the parameters in Cobb–Douglas and R is the correlation of the two variables in that equation.

The second analysis used the same equation, but interpreted x_1, labor, as the faculty of

the institution. The capital investment, x_2, was again taken as the size of the library collection (with recognition that it then stands as a surrogate for a number of institutional resources, not just the library's collections as such). The measure of productivity, x_0, for the institution is also difficult to measure, but two surrogates are used, each focussed on the research production of the faculty: (1) Ph.Ds produced and 2) publications. The sources of data were again the ARL Statistics, supplemented with data from the Social Science Citation Index for numbers of publications attributed to each university as a "corporate source". Those data were then subjected to two regression analyses, with the following results, again for 1975/76:

Labor = Faculty	$\log(a)$	$1-b$	R
x_0 = Ph.D.s produced	−0.95	0.79	0.790
x_0 = Publications	0.23	0.83	0.837

The implications of these results are clear: The larger the collection size per faculty member, the greater the research productivity (as measured by number of publications). Of course, as with all tests of this kind, it must be recognized that the variables are almost certainly not causally related. And in this case, at least, the size of collection is serving as a surrogate in itself for the whole set of capital investments made by the institutions. But the fact of such a clear relationship between the investment in the basic information resource and the productivity of faculty, whether causal or not, is important to recognize.

These results are interpreted as providing substantial confirmation that the library collection of a major university represents well the capital investment made by the institution, and that the Cobb–Douglas production model is applicable to evaluation of allocation issues in that context.

4. DISTRIBUTION OF USE

As was pointed out in the Introduction, Zipf's law does not provide an accurate description of the usage of information resource items over the entire range of data. A series of studies was conducted to test the applicability of the mixture of Poisson distributions to the usage of information related entities (such as books and index terms).

4.1 *Sources of data*

Two data bases in this study were taken from the reports of investigators at the universities of Pittsburgh and Lancaster[5, 6]. Another data base was taken from tapes of the ERIC processing center, covering postings to ERIC descriptors and identifiers for both RIE and CIJE.

4.2 *Distribution of use of library materials*

Hayes (1981) presents the results from analysis of data from the University of Pittsburgh concerning the circulation and in-house use of the collection at the Hillman Library derived by applying a "mixture of Poisson distributions" to those data. The results of the analysis suggest that circulation is NOT an adequate index of all use. The implications of that result are shown, as they apply to the issue of allocation of materials to remote storage, for comparison with a similar analysis presented in the Pittsburgh study. They show that while there is likely to be only minor effect upon "circulation usage" of the collection, there would be dramatic effect upon the "in-house usage" of the collection, with as much as 25% of that usage being adversely affected.

Beyond evaluating the validity of use of circulation data as the index to the total utilization of a collection of library materials, the paper tests the hypothesis that the proposed mixture of Poisson distributions can describe and predict the distributions of various uses. It does so by applying the mixture of Poisson distributions to eight essentially independent sets of data from the Pittsburgh Report. Taken together, the tests provide

substantial confirmations of the hypothesis, as measured by the respective coefficients of association (which vary from a low of 0.76 to a high of 0.97). The range of contexts involved, in each of which the mixture of Poisson distributions provides a qualitatively close match to the actual data, gives added weight to confidence in the applicability of that model to these kinds of data.

4.3 *Circulation use vs in-house use*

In the paper by Hayes (1979B), data from a paper by Michael K. Buckland and Anthony Hindle are analyzed to identify the extent to which alternative hypotheses match the actual circulation and in-house use of volumes, as reported by those data. Specifically, three hypotheses about the ratio, R, of in-house demand to circulation demand are considered: (1) the ratio R increases as circulation decreases; (2) the ratio R remains constant; (3) The ratio R decreases as circulation decreases.

The second of these is essentially the statement implied by Fussler–Simon and by Kent (namely, "... circulation is a valid index of all use..."); the third is that implied by Buckland and Hindle.

The basis for evaluation of these hypotheses is the fitting of mixtures of Poisson distributions to the two sets of data. The results suggest that the first hypothesis is significantly (at the 95% confidence level) more accurate in describing the actual relationship than is either of the other two.

4.4 *Distribution of use of terms from the ERIC thesaurus*

The working document, Hayes (1982C), presents results from analysis of the distribution of use of terms from the ERIC thesaurus. Data were obtained from the "postings" files for both RIE and CJIE on the frequency of assignment of each term (about 5000 in each case). These frequency distributions were then fitted with a mixture of Poisson distributions in order to test the hypothesis that the groups of terms, identifiable from the average frequency for each of the components in the mixture, would be qualitatively different.

Four components were identified. Samples of 100 terms were randomly selected from those with frequencies around the average frequency associated with each component. These sets of sample terms were then compared with respect to qualitative characteristics. Specifically, their respective positions in the thesaurus hierarchy were compared, in order to test the hypothesis that terms with greater frequency of occurrence would be at higher positions in the hierarchy. A fifth sample was selected from the entire vocabulary for purposes of comparison.

The following tabulates the statistics relating to the samples of terms and the relative statistics relating to position in the hierarchy:

Sample	1	2	3	4	5
Range of Postings	50 to 150	550 to 650	1700 to 2300	2300 to max	
Average number of levels in term hierarchy	2.88	3.02	3.33	3.35	3.00
Number of non-isolated terms	9	27	45	47	10
Number of isolated terms	16	11	6	3	9

These results confirm a general qualitative difference among the terms sampled around

the frequencies associated with each component of the mixture of Poisson distributions that best fits the distribution of frequency of postings.

5. THE EFFECT OF DISTANCE UPON THE USE OF INFORMATION RESOURCES

It has been recognized that the usage of an information system is an inverse function of distance of the user from the information resource. Of course, "distance" can be interpreted very broadly, including psychological distance as well as physical distance. Two models have been proposed to reflect the effects of distance (see Introduction), and a set of studies have been carried out to test the validity of the models.

5.1 *Source of data*

The data for the "distance study" were obtained from a survey made in 1978 of the users of the Los Angeles Public Library system, covering the Central Library and twenty branches.

5.2. *Review of the relevant literature*

First, a survey of the literature relating to the effect of distance upon the use of public libraries was undertaken. The results were reported in Palmer (1981). That paper reviews the historical studies and current trends and examines the reported data concerning specific issues, such as the mode of travel, the elasticity of demand, the subjective perception of distance, characteristics of the user, and the effects of spacing of branches. The paper then reviews some of the more general literature about the theory of location for public facilities—the "central place" theory, the "distributed goods" theory, the effects of travel, the "gravity models", the elastic demand models.

5.3 *Tests of the distance model*

The paper by Hayes (& Palmer) (1982B) presents results from examination of the relationship between distance (of the patron from the library) and the use of the library, using data from a 1978 survey of the patrons of the Los Angeles Public Library. The survey included responses from 3,662 patrons—1,068 at the Central Library and 2,594 from twenty branches. This paper presents some of the overall statistics, comparing them with the related results from a survey made ten years earlier, in 1968. It then discusses data for the entire set of branches and shows the effects of age, education, and mode of transportation used—each relating to usage as a function of distance.

Three contexts are examined as specific case studies. One is of a suburban branch (the Canoga Park branch); the second is of three closely spaced branches in a densely populated urban section; the third is of the Central Library. The effects identified from the overall statistics continue to be confirmed, but there are major differences in the areas as well as the demographic structure of the population served, resulting in major differences in the behavior. Specifically, the Canoga Park Branch exhibits heavy reliance on the automobile, with comparatively inelastic demand (i.e. slower rates of decay in use as the distance increases) from an essentially white, middle class population. The three urban branches (Felipe de Neve, Pio Pico and Memorial) serve very mixed ethnic populations and have widely varying physical settings. The Central Library serves primarily the business community of the downtown area, with a high proportion of users with at least a college education; the automobile plays a relatively lesser role as means of transportation, though, because of the parking problems in the area.

The closing of the Pio Pico branch shortly after the 1978 survey provides the opportunity to test the effects of such closings upon use. The analysis suggests that, even in an area with branches close together, the effects of distance are to cause the loss of the great bulk of the prior users of any branch that is closed. Those effects are amplified if the group that had been served is, on the average, younger or less well educated or from an ethnic minority.

In each case, usage declines as a function of distance. For the mode of transportation, the rate of decline is as predicted—exponential for walking and inverse square for driving. The effects of education are to make the rate of decline (in use as a function of distance) less as the level of education increases.

6. APPLICATION OF "WEIGHTED ENTROPY" TO INFORMATION RETRIEVAL

Two annotated bibliographies of references relevant to the study of weighted entropy for its applicability to information retrieval system design and evaluation have been developed. The first reviews the literature specific to "weighted entropy" as a measure of "significant information". The second reviews the literature specific to superimposed codes, since they provide examples of "error" which can be analyzed for the relationship between the measure, the related system design issues, and effects of error. See Hayes (1981B).

A second paper presents a preliminary discussion of the applicability of the weighted entropy model to information retrieval system evaluation and design. See Hayes (1982D).

7. SUMMARY AND CONCLUSIONS

The goals of this project were:

(1) To identify variables affecting the usage of information systems and to provide a better understanding of their roles.

(2) To provide, where possible, design equations in the form of mathematical models relating those variables, on which to base both economic and technical decisions about the structure and operation of information systems.

(3) To test and validate those models.

For the most part, the goals have been achieved, and the results will be summarized in terms of the four major models considered.

7.1 *The Cobb–Douglas model applied to industry*

The Cobb–Douglas model was applied to the investment made by manufacturing industries in the United States, with the factors of production taken as labor, capital, expenditures for information purchases, and expenditures for other purchases. The data covered the years 1967 and 1972. The results were a clear confirmation of the applicability of the Cobb–Douglas production model to the evaluation of the role of information purchases as a factor of production. Furthermore, the results imply that manufacturing industries of the United States are substantially UNDER-utilizing information as a factor in production, if they want to achieve optimum return to added value and to profit.

The work done under this project led to initiation of a related study, funded by UNESCO, to develop a measure by which the ability of countries to generate, utilize, and transfer information could be reliably estimated. The measure, called the "Information Utilization Potential (IUP)" has been applied to data from 34 separate countries.

The objectives of the study were met with respect to this model, in this context.

7.2 *The Cobb–Douglas model applied to libraries*

The Cobb-Douglas model was applied, separately, to public libraries of the several states and to major academic research libraries of the United States. The results for public libraries were very consistent, both from state to state and for the country as a whole. They demonstrate a clear relationship between public library productivity, as measured by circulation, and the investment in the collection and branch library structure of public libraries.

The application to the libraries of the Association of Research Libraries was equally clear and consistent. It demonstrates not only the importance of the collection, as a capital resource, to the productivity of the library as such, but of even greater significance, the importance of the collection of the library to the research productivity of the faculty.

Beyond the value of this application of the Cobb–Douglas model to libraries as such, there is value also in the additional confirmation of the applicability to a specific industry (libraries in this case) with such evident consistency of the parameters.

The objectives of the study with respect to this model, as it applied to libraries, were met.

7.3 *Distribution of use*

The mixture of Poisson distributions was applied to the use of books from libraries

and to the use of terms from a thesaurus. In each case, the results were a substantial confirmation of the predictive power of the model for distribution of activity as well as of the accuracy in description of the frequency distributions themselves.

The objectives of the study with respect to this model were met.

7.4 *Effect of distance on use*

The results from this part of the project were two-fold. First, there was a thorough review of the relevant literature, both as specific to location planning for libraries and as theoretically of more general application.

Second, the models of usage as a function of the cost of travel were applied to data from a survey of the users of the Los Angeles Public Library. The results confirmed the theoretical predictions, relating to the differential effects of the means of transportation. They also demonstrated the basis for differences in the effects of distance due to demographic characteristics, such as age, education, and ethnicity.

The objectives of the study with respect to this model were met.

7.5 *Applicability of weighted entropy*

Results from this aspect of the study were limited to a review of the relevant literature and to a description of the possible applications of weighted entropy to information retrieval systems and their operation. The original objectives for this model were not met.

8. PUBLICATION FROM THE PROJECT

Hayes (1979) Robert M. Hayes, The management of library resources: the balance between capital and staff in providing services. 1 March 1979. (Published in *Library Res.* **1**(2), March 1979. pp. 119–142). (Presented at AALS meeting, Jan 1980).

Hayes (1979A) Robert M. Hayes, Optimum allocation of resources: an application of the Cobb–Douglas model. 3 May 1979. (working paper).

Hayes (1979B) Robert M. Hayes, Circulation and in-house use: analysis of data from the University of Lancaster. 27 October 1979. (working paper).

Hayes (1979C) Robert M. Hayes, Mixtures of Poisson distributions (Yule's Law): a review. 1 December 1979. (working paper).

Hayes (1980) Robert M. Hayes, The information economy and national productivity. 28 February 1980. (Published in *IRCIHE Bull.*).

Hayes (1980A) Robert M. Hayes, Application of the Cobb–Douglas model to public libraries in the United States. 1 October 1980. (working paper).

Palmer (1980) James Palmer, Allocation of library resources: formulas and methods (a review of the literature). 1 December 1980. (working paper).

Borko (1981) Harold Borko, The use of resource allocation models in libraries. 3 March 1981. (Published in the *Proc. ASIS Midyear Meeting*, May 1981, Colorado; also presented at a NATO sponsored conference in Oslo, Norway on 23–30 April 1981).

Borko (1981A) Harold Borko, Information and Productivity. 21 July 1981. (Presented at the 8th Cranfield Conference on Mechanized Information Transfer, 21–24 July 1981).

Palmer (1981) E. Susan Palmer, The effect of distance upon public library: a literature survey. 5 May 1981. (Published in *Library Res.* **3**(4), Winter 1981, pp. 315–354).

Hayes (1981) Robert M. Hayes, The distribution of use of library materials: analysis of data from the University of Pittsburgh. 1 July 1981. (Published in *Library Res.* **3**(3), Fall 1981, pp. 215–260; also, presented at the ASIS National Meeting in 1981 and published in summary in the proceedings of that meeting).

Hayes (1981A) Robert M. Hayes, Weighted entropy: a review of the literature. 1 October 1981. (Working paper).

Hayes (1981B) Robert M. Hayes, Regression analysis of the Cobb–Douglas model. 28 January 1982. (Working paper).

Hayes (1982) Robert M. Hayes, Ann Pollack and Shirley Nordhaus, Application of the Cobb–Douglas model to the Assocation of Research Libraries. 1 February 1982. (To be published in *Library Res.*).

Hayes (1982A) Robert M. Hayes and Timothy S. Erickson, Added value as a function of inform. investment. 15 March 1982. (Published by the *Inform. Society* **1**(4), pp. 307–338; presented at ASIS meeting, Southern Ohio Chapter, Lexington, Ky, 30 March 1982; presented at seminar at University of Zagreb, 25 May 1982).

Hayes (1982B) Robert M. Hayes and E. Susan Palmer The effects of upon use of libraries: case studies based on a survey of users of the Los Angeles Public Library—central library and branches. 15 March 1982. (To be published by *Library Research*).

Hayes (1982C) Robert M. Hayes, The distribution of use of index terms: analysis of data from the ERIC thesaurus. 31 March 1982.

Hayes (1982D) Robert M. Hayes, The application of weighted entropy to information retrieval system evaluation & design. 31 March 1982. (working paper).

Borko (1982) Harold Borko, Information utilization: an indicator of national development. 13 June 1982. (Presented at the ASIS Midyear Meeting, Knoxville, Tennessee, 13–16 June 1982).

9. REFERENCES

[1] S. CRAWFORD, "Information needs and users. *ARIST*, Vol. 13, chap. 3, pp. 61–82. NY.: Knowledge Industry Publications, White Plains, New York. (1978).

[2] F. NARIN and J. K. MOLL, Bibliometrics. *ARIST*, Vol. 12, Chap. 2, pp. 35–58. Knowledge Industry Publications, White Plains, New York (1977).

[3] M. U. PORAT, *The information economy*. U.S. Department of Commerce, Washington, D.C. (1977).

[4] M. R. RUBIN and E. I. TAYLOR, *The* 1972 *Input–Output Study of the Information Economy*. (mimeographed publication, no date or source).

[5] A. KENT, *et al. Use of Library Materials*: *The University of Pittsburgh Study*. Marcel Dekker, New York (1979).

[6] M. K. BUCKLAND and A. HINDLE, In-house book usage in relation to circulation. *Collection Management*. 1978 **2**(4), 265–278.