CHAPTER 4

FORMULATION OF THE JOURNAL SELECTION PROBLEM

4.1 Introduction

The problem we consider is that of allocating a serials budget for one decision period (which may equal one or more subscription periods) in such a way that the total expected document exposure time purchased is a maximum. The use of bi-valent decision variables is appropriate for such a problem. Hence we define variables X_j, where X_j takes the value unity if the jth journal is purchased, and is zero otherwise.

It is assumed that the equitable distribution of resources among the M departments is part of university policy, and so we consider constraints of the general form;

$$P_i \geq L_i ; \qquad i = 1, \ldots, M, \qquad (1)$$

in addition to the prescribed budget constraints. Here, P_i is the expected document exposure for the ith department, and L_i is the minimum document exposure that the ith department should, in fairness, expect.

4.2 The Objective Function

As discussed in Section 3.2, any set of acquired volumes of primary periodicals will have a certain potential exposure, the realization of which depends on a number of factors, including:

(1) Availability and effectiveness of secondary periodicals;

(2) Accessibility of periodicals stocks;

(3) Condition of periodicals stocks.

These factors are all subject to some control by the librarian, who must make decisions regarding the selection of secondary periodicals, and the storage, discarding, and binding of primary periodicals.

Since the aim is to maximize expected document exposure, the selection of a primary periodical will be influenced by the following:

(1) Whether or not it will receive coverage in the secondary periodicals held by the library;

(2) How long it should, or can, remain on open access;

(3) When, if ever, it will be bound.

Conversely, decision making in these three areas must be influenced by the set of primary periodicals selected for subscription.

Ideally, then, the selection of primary periodicals should be treated as part of a wider decision problem which includes the selection of secondary periodicals, and the determination of storage, discarding, and binding policies.

Although it might be possible, in principle, to formulate an analytical approach to this decision problem, the practical difficulties are enormous, as briefly illustrated below.

(a) Secondary Periodicals

Suppose there are S secondary periodicals, and define bivalent decision variables Y_k for their selection. Then C_j, the expected document exposure of the jth primary periodical, must be a function of the variables Y_k :

$$C_j = C_j (Y_1, \ldots, Y_s), \qquad (2)$$

and the objective function is then a sum of terms of the form

$$X_j C_j (Y_1, \ldots, Y_s). \qquad (3)$$

That is, the objective function is non-linear in the decision variables. If the function $C_j (Y_1, \ldots, Y_s)$ takes the particularly simple form of a linear combination of the variables,

$$C_j = r_j + \sum_{k=1}^{s} r_{jk} Y_k, \qquad (4)$$

the problem presented by this non-linearity can be handled, at least in principle. However, in a problem which is already very large, the extra dimension introduced by the non-linearity would almost certainly make solution a practical impossibility. In any case, the simple form (4) for C_j will not hold in general, because heavy overlap usually exists among the coverages provided by different secondary periodicals, and so the r_{jk} are not really defined quantities - the incremental contributions of a secondary periodical depends on which secondary periodicals have already been accepted.

It appears, therefore, that the selection of primary and secondary periodicals will have to be treated as two separate (but interacting) decision problems, each with a prescribed budget.

Although no models of the process exist, it seems reasonable to assume that the effectiveness of the acquisition of secondary periodicals in increasing document exposure is directly related to the degree of coverage of all candidate

periodicals that is achieved by that acquisition. If we accept this principle, the selection problem for secondary periodicals might be formulated as that of allocating a prescribed budget so that the proportion of candidate journals covered is a maximum, subject to equity constraints.

An algorithm for the solution of this problem has not been found.

(b) Storage and Discarding Policies

The expected document exposure of the jth primary journal must also be a function of the decision variables T_j^c, the age at which a volume of the journal is moved into closed-access storage, and T_j^d, the age at which a volume is discarded altogether. The objective function is then a sum of terms of the form,

$$X_j \; C_j \; (T_j^c, \; T_j^d). \qquad (5)$$

The form taken by this functional dependence is not known with certainty. However considerable effort has been put into the development of policies for discarding material and for transferring items to closed stacks. The assumption is generally made that usage decays exponentially with time, and we have already pointed to the severe criticism, by Line and Sandison, of such models.

If the C_j were simple linear functions, the problem might be feasible; however there is no support in the literature for such an assumption. More complicated functional dependence

could be handled, at least in principle, by the use of piece-wise linear approximations. However this would present extreme dimensionality problems.

Thus, for reasons of practicality, we conclude that journal selection is to be considered separately from the determination of discarding and closed-access policies.

(c) Binding

The binding of journal volumes makes it easier for users to locate desired articles on the shelves, and might also reduce losses due to mutilation and theft. However there is a negative effect due to the absence of journals during the binding period. Hence the expected document exposure of the jth journal is a function of the age T_j^b at which a volume of that journal is to be bound. The objective function is then a sum of terms of the form

$$X_j \quad C_j \quad (T_j^b). \qquad (6)$$

There is no indication in the literature of the form which this function might take, although it seems reasonably clear that it will not be linear. The only reported attempt (Buckland and Woodburn, 1968) to develop a procedure for determining optimal serial binding policies assumes that all titles are to be bound, at the same age. Clearly, not all journals are worth binding, while some require binding more urgently than others. However considerable work is still required, to provide a quantitative assessment of the contribution that binding of journals can make to document exposure.

Under these circumstances, there appears at present to be

no alternative to considering journal selection and the determination of a binding policy as separate decision problems.

Thus, the selection of primary periodicals is to be considered separately from but in conjunction with, those decisions which determine the availability, accessibility and condition of journal stocks.

We concluded earlier that the exposure realization factor, f, can be expected to vary among departments, and it was also noted that no models are available for estimating its values. However the equitable distribution of library expenditure and effort should work towards reducing the variation of this factor among departments. It therefore seems reasonable to adopt the simplest possible approach, and assume that the exposure realization factor is in fact the same for all departments.

Under this assumption, the journal selection policy should simply be to maximize the potential document exposure of the set of volumes acquired. That is, one seeks to maximize the product of n and e, where n is the number of relevant articles acquired, and e is the average exposure time per article. Since e is a constant parameter, the problem reduces to that of maximizing n.

Therefore the objective function of the mathematical programming problem is

$$Z = \sum_{j \in Q_p} N_j X_j \qquad (7)$$

where N_j is the number of relevant articles contained in the jth journal, and Q_p is the set of candidate primary journals.

4.3 Equity Constraints

Considering the equitable distribution of resources, we note that P_i, the expected document exposure time for the ith department is

$$P_i = fe \sum_{j \in Q_p} n_{ij} X_j,$$

where n_{ij} is the number of articles relevant to the interests of the ith department, published in the jth journal. Note that $\sum_i n_{ij} \geq N_j$

The maximum exposure that the ith department could expect under any solution is

$$P_i \text{ (max)} = fe \sum_{j \in Q_p} n_{ij}.$$

It is proposed that an appropriate definition for L_i, the minimum exposure time that the ith department can expect, is such that

$$\frac{L_i}{P_i \text{ (max)}} = \text{a constant, } q.$$

Condition (1) of Section 4.1 then becomes

$$fe \sum_{j \in Q_p} n_{ij} X_j \geq qfe \sum_{j \in Q_p} n_{ij},$$

or

$$\sum_{j \in Q_p} n_{ij} X_j \geq q \sum_{j \in Q_p} n_{ij}, \quad i = 1, \ldots, M. \quad (8)$$

4.4 Inter-Library Loans

The connection between journal selection and the operation of inter-library-loans requires special comment.

Referring to Figure 2.5, we note that the serials selection policy helps to determine the rate and content of inter-library loan requests. Clearly, the higher the level of demand that can be satisfied directly, the lower the rate of indirect satisfaction required via inter-library loans.

In considering the overall performance measure of document exposure, it is not in general strictly possible to isolate the inter-library loans operation from the serials acquisition subsystem. It may be advantageous, for example, not to subscribe to a journal which is always readily available on loan from another library. On the other hand, significant journals which are not readily available through inter-library-loans could be more important candidates for subscription.

One might be tempted to take the narrow economic view, so that a title is purchased if its expected use for the period during which it will be stored by the library is high enough so that inter-library borrowing costs to satisfy those requests are greater than the sum of the purchase cost and the storage cost. While this approach might minimize costs for the library, it will not necessarily maximize document exposure, since the realized demand for articles in a given journal will be lower if the journal is not held by the library, because more effort is required on the part of the user, and longer delays are anticipated.

The problem might be approached as follows:

The total potential exposure of all candidate journals could be written:

$$Z = \sum_{j \in Q_p} C_j X_j + \sum_{j \in Q_p} C'_j (1 - X_j),$$

where C_j is the potential exposure of the jth journal, when that journal is held by the library, and C'_j is the potential exposure when the journal is not held.

In general, $C'_j \leq C_j$, the equality applying only when, for example, the journal is held by another local library from which it can be obtained with an absolute minimum of effort and delay.

Similarly, the expected document exposure could be written in the form

$$F = f \sum_{j \in Q_p} C_j X_j + f' \sum_{j \in Q_p} C'_j (1 - X_j),$$

where f is the exposure realization rate for journals that are held;

f' is the exposure realization rate for journals that are not held.

Serious difficulties must be overcome, before further progress can be made.

For example, we wrote, earlier,

$$C_j = eN_j,$$

where e is a constant. However in the case of C'_j, there is an additional factor, less than unity, reflecting the reduced potential of a journal that is not held by the library. Further, this factor can be expected to vary from journal to journal, because some

journals may be more readily available than others.

Also, the factors f and f' are not constant parameters, but are themselves influenced by the results of decisions concerning the acquisition of periodicals, and the budget allocated to inter-library-loans operations. That is, if an analytical formulation could be achieved, the factors f and f' would be functions of decision variables, so that practical difficulties arising from the size of the problem could make an analytical solution impossible.

Considerable effort is still required to clarify this area of library management.

At the present time, the only practical approach seems to be that of considering the journal selection problem separately from the operation of inter-library-loans, and formulating an empirical strategy for the distribution of resources between the two areas.

4.5 Subscription, Processing and the Total Budget

Costs which are directly related to the selection of primary journals are:

(1) Subscription costs (including the costs associated with obtaining and despatching material for exchange);

(2) Processing costs associated with:

booking-in new issues as they arrive;

handling payments of subscriptions;

effecting cancellations, and initiation of new subscriptions;

cataloguing new titles;

provision of information (e.g., inputs to data processing; index information).

Subscription costs naturally vary with title. Booking-in costs per issue can be assumed to be the same for all titles, so the booking-in costs for the jth title are proportional to s_j, the number of issues appearing during the decision period. The remaining processing costs for each title may reasonably be assumed to be the same for all titles.

We define the following parameters:

a_b = cost associated with booking-in one issue;

a_h = cost associated with handling the subscription payment for one title;

a_c = cost associated with effecting the cancellation of one title (correspondence; index amendment; display alterations);

a_s = cost associated with initiating a new subscription (correspondence; index amendment; display alterations);

$a_{cat.}$ = cost associated with cataloguing each new title;

a_i = cost per title associated with provision of information.

The candidate set Q_p may be divided into two parts:

$Q_p(H)$: the set of candidate primary journals that are already held;

$Q_p(N)$: the set of candidate primary journals that are not already held.

The total cost associated with any solution is given by

$$\sum_{j\in Q_p(H)} \{ X_j (a_j^{sub} + a_b s_j + a_h + a_i) + (1-X_j)a_c \} + \sum_{j\in Q_p(N)} X_j (a_j^{sub} + a_b s_j + a_h + a_s + a_{cat.} + a_i)$$

and this total cost is constrained to be less than or equal to the total prescribed budget B'. The final form of this constraint becomes

$$\sum_{j\in Q_p} a_j X_j \leq B, \qquad (9)$$

where $a_j = a_j^{sub} + a_b s_j + a_h + a_i - a_c$, for $j\in Q_p(H)$:

$a_j = a_j^{sub} + a_b s_j + a_h + a_s + a_{cat.} + a_i$, for $j\in Q_p(N)$,

$B = B' - N_p(H)a_c$,

$N_p(H)$ = the number of candidate journals already held.

Output from a solution of the journal selection problem, for a given total budget, will provide the distribution of that total budget among the various activities involved.

Our formulation of the journal selection problem is therefore:

Maximize $Z = \sum_{j\in Q_p} N_j X_j$,

Subject to $\sum_{j\in Q_p} a_j X_j \leq B$,

$$\sum_{j\in Q_p} n_{ij} X_j \geq q \sum_{j\in Q_p} n_{ij} ,$$

$X_j = 0, 1.$

4.6 Data Requirements

Data requirements for this model are severe. The number of data elements needed depends mainly on the number of candidate journals that are relevant to the interests of each department individually. Thus, of a total candidate set of, say, 10,000 journals, being considered for fifteen departments, the average number that are of interest to a department is not likely to exceed 1000, so that about 15,000 non-zero values of n_{ij} may be required. While not beyond the bounds of feasibility, this does present a major data-gathering problem.

The most important data required are those relating to subscription costs (a_j^{sub}), and the numbers n_{ij} of articles relevant to the interests of the ith department, and published in the jth journal.

These quantities are not available for direct measurement:

(a) because it is generally not possible to obtain firm quotes of subscription prices before placing a subscription;

(b) because variations in currency exchange rates can cause significant changes in the actual subscription costs incurred by the library (this is particularly so for Australian libraries, since most journals are published overseas);

(c) because obviously the selection decision must be made before the relevant articles are actually published, i.e., before the numbers n_{ij} are even defined.

Thus the determination of these data is a matter of extrapolation on the basis of already established data. Other data

might be obtained by direct costing studies, or by referring to library standards (Robertson and Hensman, 1975).

4.6.1 Subscription Costs

The 'established data' for subscription costs which might be available would relate only to those titles already received. Even these data will be difficult to extract from accounts of payment, unless the data processing system is explicitly designed to provide it, as anyone acquainted with the cash flow problems of library acquisitions would readily confirm.

Subscription agents (e.g., EBSCO) usually provide pre-publication lists of expected subscription prices, based on current information from publishers. It would appear that these lists provide the only practical source of data and that one must accept the likelihood of unpredictable errors. The data must be standardized by expressing them all in the same currency, using exchange rates prevailing at some prescribed date. A further problem is that there may be some candidate journals which are not covered in these lists. Price estimates may then have to be obtained by direct negotiation with publishers.

It is reasonable to assume that actual costs will be, on the average, higher than those indicated by the above sources. On the other hand, the budgets that will actually be realized are not guaranteed to be higher than the levels used in determining policies. Therefore the cost data should be inflated by some factor which is the same for all titles. Regular price index studies such as that of Brown (1976), cannot be used directly to provide this factor,

since they relate to variations in actual subscription prices from year to year (i.e., the prices information actually provided in the journal issues themselves).

The 'inflation' factor might be estimated by a study which compares journal price information as provided in the issues of journals taken in the current year, with the prices indicated in the previous year by the sources mentioned above. (If the decision period is longer than one year, then similar studies could be made to provide 'lumped' estimates.) Standard sampling techniques could be used to determine the sample size necessary to give a good estimate.

Individual journal subscription costs can still be expected to vary, in a random manner, from the estimates obtained and used in the solution of the selection problem. Therefore tests will be required in order to evaluate the sensitivity of the solution to such variation, and to determine the necessity, or otherwise, for more strenuous efforts to obtain better data.

4.6.2 Productivities

The problem of defining the potential demand for serials literature reduces, in the current model, to that of finding values for the productivities n_{ij}.

The determination of these productivities, and their use in journal selection, is an old topic, dating back at least to the work of Bradford (1948). Consequently there is a considerable amount of literature devoted to the subject. Some of this work (e.g., Brookes, 1968; Leimkuhler, 1967) is concerned with the distribution of productivity among journals, and a theoretical understanding of how

these distributions are generated. However the bulk of this literature is devoted to the ranking of journals, for the purpose of selection, on the basis of productivity (e.g., abstract counts), measured library usage, or citation ratings.

Citation analysis is a method of retrieving articles of relevance to a given subject area. The subject area is defined by a small set of well known 'key' articles, each of which provides the starting node or root, of a citation tree which is constructed by branching from a given article to each of the references cited by that article. Journals are ranked according to the number of cited articles they produce. Such rankings have tremendous appeal for selection purposes because they apparently reflect the actual usage of journals. Also the development of large data bases such as that compiled for the production of Science Citation Index is making automatic citation analysis possible.

The results of citation analysis so far have been disappointing. A study which compared usage ratings at the N.L.L. with citation rankings of the same set of journals showed a very low correlation between the two rankings (Scales, 1976). To quote Scales: "These results seem to show that journal citation rankings . . . . . do not constitute valid guides for journal selection.".

A review of the possible biases involved in citation analysis is beyond the scope of the present work (see Scales, 1976, for some examples). However we have adopted a model in which expected document exposure is related to journal productivity by some as yet undeveloped model of user behaviour. Part of this user behaviour would be a preference for so-called 'prestigious' journals, and if

some of the bias can be removed from citation analysis it will prove to be a valuable research tool in this area. For example, the journal 'impact' factor (Garfield, 1972) defined as the number of citations received by the journal, divided by the number of articles published in the journal, could provide a basis for weighting a journal's productivity in accordance with its prestige.

Productivity measurements are made by searching a subject index of published articles, according to some prescribed subject interest profile. The retrieval of information in this way has been the subject of concentrated attention for many years, and the range of data bases that have been compiled for automated searching has become quite large (Vickers, 1973; Leggate, 1975; Dammers, 1975; Smith, 1976). So far as journal selection is concerned, the degree to which the total literature is covered by indexing and abstracting data bases could still be an important limitation and is a subject for further study.

In cases of incomplete coverage, one must rely on individual bibliographies compiled over time by the library users.

Processors of data bases provide output in three classes. As new material is added to up-date the data-base, it can be searched to provide current-awareness 'alerts' either as a personalized service to individual subscribers (SDI - Selective Dissemination of Information), or as a broader coverage in accordance with group profiles of organizations. Thirdly, retrospective searches of the data bases can be made as 'one-off' operations.

The main organizational options are:

(1) Subscribe direct to commercially available searches, conducted and retailed either by the original compilers of the data base, or by leasing organizations.

(2) Lease data-base tapes from the organizations, and conduct in-house processing for members of the organization.

Vickers (1973) provided an analysis of costs incurred by organizations operating under different options. Because of the very wide range of subjects involved in the academic library system, option (2) would be precluded by the high level of fixed costs (leasing of tapes; processing establishment).

A complete survey of available subscription services has not been possible, but Vickers' (1973, p. 273) table of the variable costs (to the supplier) of retrospective searches may be indicative of the scale of costs for subscription services. These range from a high of $133.70 (U.S. dollars) for a search of a file of 1,000,000 items, to a low of $3.30 to search a file of 40,000 items. These costs cover search formulation, computer processing, output printing and reproduction, output checking, distribution and telecommunication. A common unit for the assessment of computer processing costs is the cost per 'search term' (the search profile is a list of terms representing the user's interests in the data base, and a statement of the logical conditions required for a match). Vickers gives figures ranging from 17 cents to 107 cents per search term, for retrospective searches.

Although retrospective searching is an actual or intended part of the function of these information broking systems, SDI is still

the main product (Leggate, 1975). As an indicator of the scale of operations in this area, the Georgia Information Dissemination Centre, in 1972, was running 3,500 SDI profiles against 11 (leased) data bases. According to Leggate, most SDI systems form part of a much larger computer, information, library or publication operation, and SDI costs have to be disentangled from the costs of the complete system. Vickers (1973) found total variable costs per item output ranging from 5 cents to 111 cents, while Claasen (reported in Dammers, 1975) found composite annual processing costs per profile term of around £1 for a data base such as Chemical Abstracts Condensates.

A tabulation of subscription services in Australia was given by Smith (1976), indicating some monthly services still available at no charge, with the most expensive being a weekly service against Chemical Abstracts Condensates provided by CSIRO at $100 - $200 p.a. More recent information published in Australian Special Libraries, Vol. 9, indicates SDI services provided by the ANL at $50 p.a., and retrospective searches of MEDLARS costing $20 for 1971 - 1976 and $15 for 1966 - 1971. CSIRO also provides retrospective searches at comparable prices (J. Tonnoir, personal communication), but outputs are limited, for example, to 200 items.

Unit costs are decreasing, mainly as a result of the rapid decline in computer processing costs (Leggate, 1975), although the component due to the intellectual effort required for search formulation will act to dampen the rate of decrease.

A recent enquiry in Australia recommended the establishment of large central data bases (at the National Library) from which SDI and retrospective search services should be made available as cheaply

as possible to users throughout Australia. It therefore seems reasonable to assume that future years will see increasing availability and use of current-awareness services, at least in scientific areas (both social and technological).

Such expansion would have tremendous impact on the data - collection problems we are considering here, since it may be possible to obtain sufficiently accurate productivity measures for many candidate journals by aggregating and analyzing the accumulated SDI bibliographies of individual users for, say, the preceding three years. This automatic data collection aspect of SDI services represents a further argument in favour of their establishment.

CHAPTER 5

SOLUTION ALGORITHMS AND A SAMPLE PROBLEM

5.1 Introduction

The journal selection problem formulated in Chapter 4 is presented here for convenience:

$$\text{Maximize} \quad Z = \sum_{j \in Q_p} N_j X_j \,,$$

$$\text{Subject to} \quad \sum_{j \in Q_p} a_j X_j \leq B \,, \tag{10}$$

$$\sum_{j \in Q_p} n_{ij} X_j \geq L_i = q \sum_{j \in Q_p} n_{ij}; \quad i = 1, \ldots, M,$$

$$X_j = 0, 1.$$

This is a single-knapsack problem in N zero-one variables, with M demand-satisfaction constraints. The size of the "knapsack" is represented by the budget, B, while the items selected to be carried in the knapsack, for maximum value, are represented by the journals selected, for maximum document exposure.

The N_j and n_{ij} are integers, with values ranging from 1 to several hundred. The a_j are costs which, expressed to the nearest dollar, yield integers ranging from about 5 to several hundred, and B is an integer with a value of several tens of thousands.

Since q satisfies $0 \leq q \leq 1$, the right-hand sides of the constraints are, for convenience, rounded to the nearest integer.

The main feature of the problem is its size: N will generally be of the order of several thousand. This makes it doubtful if the conventional techniques of branch and bound, or implicit enumeration, will be capable of providing a solution within acceptable computation times.

A small sample problem in 150 variables with two demand constraints was generated, to have the main structural features to be expected of the problem (see below). A locally available coded version of a modified Balas algorithm (Petersen, 1967) was applied to the sample, and failed to converge in 2000 iterations. Storage requirements for this code are also rather excessive: a problem involving 2000 variables and five demand constraints would require 42K words. (These requirements could almost certainly be reduced by more efficient coding.)

The main structural feature of the problem is the fact that productivities can be expected to conform, in some measure, with the Bradford distribution. That is, there will be a 'core' of highly productive journals and many other journals whose contributions are very low. Thus a ranking of journals according to their benefit/cost ratios will produce a long tail of journals with very low values of this ratio. This suggests the possibility that good sub-optimal solutions might be found quite quickly by an heuristic algorithm based on the elimination of variables with low benefit/cost ratios. Note that this approach is in line with approximate methods suggested in the literature (e.g., Robertson and Hensman, 1975). The chief new requirement here is for a method of dealing with the multiplicity of departments, via the equity constraints.

An evaluation of the effectiveness of such an algorithm requires some measure of how close to optimality its solutions are. In the present case, of course, no optimal solutions of the complete problem are available for comparison. However the simple knapsack problem obtained by relaxing the equity constraints is amenable to efficient solution techniques, and provides a standard against which the heuristic solution of the full problem can be evaluated.

5.2 The Simple Knapsack Problem

In a recent paper, Nauss (1976), building on the work of Ingargiola and Korsh (1973) and Horowitz and Sahni (1974), has presented an efficient algorithm for the zero-one simple knapsack problem. For the test problems used, the new algorithm is considerably more efficient than the well-known Greenberg and Hegerich (1970) branch and bound algorithm.

Nauss' approach is in two stages. The first stage identifies those variables which must have a value of 1, and those which must have a value of 0, in any optimal solution. These variables are then effectively eliminated from the problem, and the reduced problem, in the remaining variables, is solved by a branch and bound algorithm.

Ingargiola and Korsh (1973) developed inexpensive tests whose execution is based on the fact that, once the variables are arranged in decreasing order of benefit/cost, the solution of the continuous problem obtained by replacing the condition $X_j = 0, 1$ by the condition $0 \leq X_j \leq 1$, becomes analytic. These tests allow variables to be fixed to 0 or 1, in the following manner.

We consider the problem:

$$\text{Maximize} \quad \sum_{j=1}^{N} c_j X_j \ ,$$

$$\text{Subject to} \quad \sum_{j=1}^{N} a_j X_j \leq B \ , \tag{11}$$

$$X_j = 0, 1.$$

Consider two disjoint subsets I_1, I_0 of $A = \{1, 2, \ldots, N\}$, such that $I_1 = \{j : X_j = 1\}$; $I_0 = \{j : X_j = 0\}$. For given I_1, I_0, formulate the reduced problem:

$$\text{Maximize} \quad \sum_{j \notin I_0 \cup I_1} c_j X_j + \sum_{j \in I_1} c_j \ ,$$

$$\text{Subject to} \quad \sum_{j \notin I_0 \cup I_1} a_j X_j \leq B - \sum_{j \in I_1} a_j \ , \tag{12}$$

$$0 \leq X_j \leq 1.$$

Let $C(I_1, I_0)$ be the value of the solution of the reduced problem (12). If we have a solution $Y_1, Y_2, \ldots, Y_N$ to the problem (11), with $C_0 = \sum_{j=1}^{N} c_j Y_j$, then if $C_0 > C(I_1, I_0)$, for some choice of I_1, I_0, no optimal solution $Z_1, Z_2, \ldots, Z_N$ of (11) will have $Z_j = 1$ for all $j \in I_1$ and $Z_j = 0$ for all $j \in I_0$. In particular, if $I_1 = \{j_0\}$, $I_0 = \emptyset$, and $C_0 > C(I_1, I_0)$, then X_{j_0} must be 0 in any optimal solution to (11). Furthermore, if $C_0 > C(\emptyset, \{j_0\})$, then X_{j_0} must be 1 in any optimal solution to (11).

The reduction algorithm developed by Korsh and Ingargiola is very efficient as it uses space and time proportional to N. However Nauss has developed a slightly different algorithm for which

computation time is about two-thirds less. He uses a pegging test based on the concept of Lagrangean relaxation, which although weaker than the test of Korsh and Ingargiola, is cheaper to perform.

The reduced problem is solved by a branch and bound procedure based on an algorithm of Horowitz and Sahni (1974), but with a slight alteration which is claimed to decrease computation time for this stage by about one-third.

5.2.1 Simple Knapsack Algorithm

The complete algorithm, as given by Nauss, is presented here for convenience. The problem (11) is denoted by (P), and its relaxed form (i.e., with $X_j = 0,1$ replaced by $0 \leq X_j \leq 1$) is denoted by $(\bar{P})$.

1. Order the variables in decreasing benefit/cost ratios so that

$$\frac{c_1}{a_1} \geq \frac{c_2}{a_2} \geq \cdots \geq \frac{c_N}{a_N} .$$

   Set $I_1 = I_0 = \emptyset$.

2. Solve $(\bar{P})$, getting an optimal solution $\bar{X}$ with value $v(\bar{P})$, and a shadow price p associated with the budget constraint. If $\bar{X}$ is feasible for (P), stop: the solution is optimal. Otherwise denote the index of the fractional variable by r.

3. Find a lower bound Z* for v(P) by setting $\bar{X}_r = 0$ in the solution to $(\bar{P})$. Let $X^* = \bar{X}$. Try to improve Z* by certain heuristics. (See Note (3) below.)

4. For all $i = 1,\ldots,r-1$, if $v(\bar{P}) - c_i + pa_i \leq Z^*$, set $I_1 = I_1 U\ i$ (X_i is pegged to 1).

5. For all $i = r+1,\ldots,N$, if $v(\bar{P}) + c_i - pa_i \leq Z^*$, set $I_0 = I_0 \cup i$ (X_i is pegged to 0).

6. Solve the remaining knapsack problem (12) by using the branch and bound procedure in steps 7 - 17.

7. Initialize the candidate list to consist of the reduced problem, and let the incumbent value be Z^*.

8. If the candidate list is empty, stop: X^* is an optimal solution to (P) and Z^* is the optimal value.

9. Select a candidate problem (CP) from the list by a LIFO rule.

10. Solve $(C\bar{P})$, getting an optimal solution $\bar{X}$.

11. If $(C\bar{P})$ is infeasible, go to 8.

12. If $v(C\bar{P}) \leq Z^*$ go to 8.

13. If an optimal solution of $(C\bar{P})$ is feasible in (CP), go to 17.

14. Choose that X_j which is the free variable with the largest benefit/cost.

15. If $a_j \leq B - \sum_{j:X_j \text{ set to } 1} a_j$, then add only $(CP : X_j = 0)$ to the list, add the restriction $X_j = 1$ to (CP), and go to 14. Otherwise, go to 16.

16. If $a_j > B - \sum_{j:X_j \text{ set to } 1} a_j$, add the restriction $X_j = 0$ to (CP), choose the X_j with the largest benefit/cost, and return to the beginning of this step. Otherwise, go to 10.

17. A feasible solution to (P) has been found. Set $Z^* = v(C\bar{P})$, $X^* = \bar{X}$, and go to 8.

Notes: (1) In steps 2. and 10., solution of the problems is analytic:

$$X_j = 1 \text{ for } 1 \leqq j < r \; ; \quad j \notin I_1 \cup I_0 \; ;$$

$$X_j = 0 \text{ for } r < j \leqq N \; ; \quad j \notin I_1 \cup I_0 \; ;$$

$$X_r = \Big(B - \sum_{j \in I_1} a_j - \sum_{j \notin I_1 \cup I_0;\ j < r} a_j \Big)/a_r \; ;$$

with r defined as the smallest $k \notin I_1 \cup I_0$ such that

$$\sum_{j \notin I_1 \cup I_0 \,:\, j \leqq k} a_j \geqq B - \sum_{j \in I_1} a_j \; .$$

(2) The shadow price p is easily shown to be c_r/a_r. It can also be shown (Hadley, 1964, p. 73) that the shadow price is also the optimal Lagrange multiplier for $(\bar{P})$.

(3) Heuristics involved in step 3. are as follows: First, set $\overset{''}{X} = \bar{X}$ and $\overset{''}{X}_r = 0$. The solution $\overset{''}{X}$ then has a slack in the constraint with value $s = \bar{X}_r a_r$. Now, for $i = r + 1, \ldots, N$, the following is done: if $a_i \leqq s$, set $\overset{''}{X}_i = 1$ and $s = s - a_i$. If $s > 0$, repeat this step for $i = i + 1$. If $c\overset{''}{X} > cX^*$, set $Z^* = c\overset{''}{X}$ and $X^* = \overset{''}{X}$. This heuristic puts extra variables in the knapsack until no more fit.

The second heuristic begins by setting $\overset{''}{X} = \bar{X}$ and $\overset{''}{X}_r = 1$. This overfills the knapsack by $s = (1 - \bar{X}_r)a_r$. Then for $i = r - 1, r - 2, \ldots, 1$, the following is done: set $\overset{''}{X}_i = 0$, $s = s - a_i$, and if $s > 0$ repeat this step for $i = i - 1$. When $s \leqq 0$, set $s = -s$ and return to the test loop in the first heuristic. This heuristic

begins by overfilling the knapsack, and then variables are withdrawn until feasibility is obtained.

The algorithm (KNAP1) was coded in FORTRAN and tested on 50 problems each of 20 variables. In the first set, of 25 problems, the c_j were generated randomly from a uniform distribution over the range (10,50), while the a_j were similarly obtained from the interval (1,20). In all cases, the budget was set at $.5 \sum_{j=1}^{N} a_j$. For the second set of 25 problems, the intervals were changed to (10,100) and (1,50) respectively.

All of these problems were also solved using the locally available coded version of Balas' algorithm (Petersen, 1967), to provide an accuracy check. For problems of this size and structure, performance of the new algorithm was not significantly better than the Balas code.

The storage and retrieval of candidate problems presents a potential difficulty for branch and bound algorithms. However in the present case, each candidate problem can be completely defined by the list of variables set to unity or zero. Hence each candidate problem is stored in the form of a bit-string. This is accomplished by the use of byte-packing and -extracting MACRO subroutines (for which I am indebted to A. Griffiths). It should be noted that no attempts have been made at this stage to maximize coding efficiency.

5.3 Knapsack Problem with Demand Constraints

Here, we consider the following problem:

$$\text{Maximize} \quad Z = \sum_{j=1}^{N} c_j X_j ,$$

Subject to

$$\sum_{j=1}^{N} a_j X_j \leq B ,$$

$$\sum_{j=1}^{N} n_{ij} X_j \geq L_i \; ; \quad i = 1, \ldots, M \; ; \qquad (14)$$

$$X_j = 0,1.$$

It is a relatively straightforward matter to devise a simple algorithm which finds a good feasible solution to this problem. In the approach used here, variables are rejected from the solution, in increasing order of benefit/cost, until the budget constraint can be satisfied. However, before each variable is dropped, it is examined to see if its rejection will cause any of the demand constraints to be violated; if so, the variable is retained. Once feasibility is attained, the rejected variables are again checked, in decreasing order of benefit/cost, and restored to the solution whenever this can be done without violating the budget constraint.

As demonstrated below, this simple algorithm (JNLSEL) can very quickly find an extremely good solution to a problem with some of the main structural features of the journal selection problem. If feasibility cannot be reached because the budget is too low in relation to the equity level, the algorithm reaches a low-cost solution which satisfies the prescribed equity constraints.

However it should be noted that the algorithm can fail to find a feasible solution, when in fact one exists. Attempts have not been made to remedy this short-coming, and in fact it is not clear that the extra computational effort would be worthwhile, since the <u>absolute</u> value of the equity factor q (Section 4.3) is not likely

to be critically defined. As would be expected, the results obtained so far demonstrate that as q is increased, for a fixed budget, the range of variation among the quantities $R_i = \sum_{j=1}^{N} n_{ij} X_j$; $i = 1, \ldots, M$, decreases (i.e., the solution becomes more equitable.). At the same time, the overall value of the solution, Z, decreases. That is, a trade-off occurs between equity and total return, so the choice of solution (i.e., a choice of a value for q) is really a matter for value-judgment by the librarian.

5.4 A Sample Problem

The collection of real data in order to evaluate the approach described so far, is beyond the scope of the present work. However progress can be made in testing the algorithm against artificially generated data intended to reflect the main structural features of a real data set.

5.4.1 Productivities

The chief aid in the generation of these data elements is the Bradford - Zipf distribution (Section 3.3). The representation of Brookes (1969) was chosen, in which $R(n)$, the cumulative sum of articles in the n most productive journals, is given by

$$R(n) = \alpha n^{\beta} \quad ; \quad 1 \leq n \leq c \,,$$

$$R(n) = k \log_e \left(\frac{n}{s}\right) \quad ; \quad c \leq n \leq N \,, \qquad (16)$$

where: α, β, k, s and c are parameters characteristic of the bibliography being considered:

α is the number of articles in the most productive journal,

β is a constant less than 1;

k is a constant whose value may be quite large (e.g., 1000);

s is a parameter which is related to the breadth of the subject area being considered. Generally, $s \geq 1$, and $s = 1$ for very narrow subjects;

c is related to s, and has not yet been found to be less than 3. It defines a 'nuclear zone' of highly productive journals.

A graphical representation of the distribution is presented in Figure 5.1.

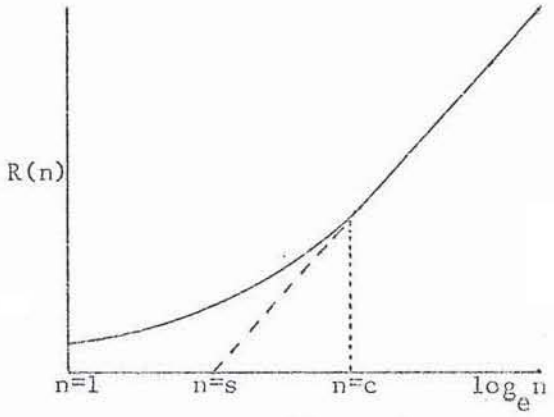

Fig. 5.1

The Bradford Distribution

The parameters are connected by relations which must be satisfied for continuity at $n = c$. These are:

$$c^{\beta} = \frac{k}{\alpha \beta}; \qquad (13)$$

$$\frac{1}{\beta} = \log_e \left(\frac{c}{s}\right).$$

Eight university departments were considered in the sample problem, for which the following parameters were arbitrarily chosen (to satisfy (13)):

TABLE 5.1

	N	C	α	β	k	s
1	80	15	51	.757	300	4
2	200	50	71	.621	500	10
3	150	15	43	.757	250	4
4	120	25	45	.701	300	6
5	200	40	98	.621	600	8
6	350	30	97	.621	500	6
7	400	55	98	.657	900	12
8	100	20	29	.721	180	5

The eight distributions are illustrated in Figure 5.2.

An important feature of the problem is, of course, the overlap between subject areas. Thus certain journals are of interest to more than one department. Overlap was again arbitrarily specified, and is represented in the following matrix:

TABLE 5.2

	1	2	3	4	5	6	7	8
1		0	0	5	0	0	0	0
2			5	20	15	5	5	5
3				30	15	15	5	25
4					0	0	10	20
5						20	0	0
7							80	0
8								0

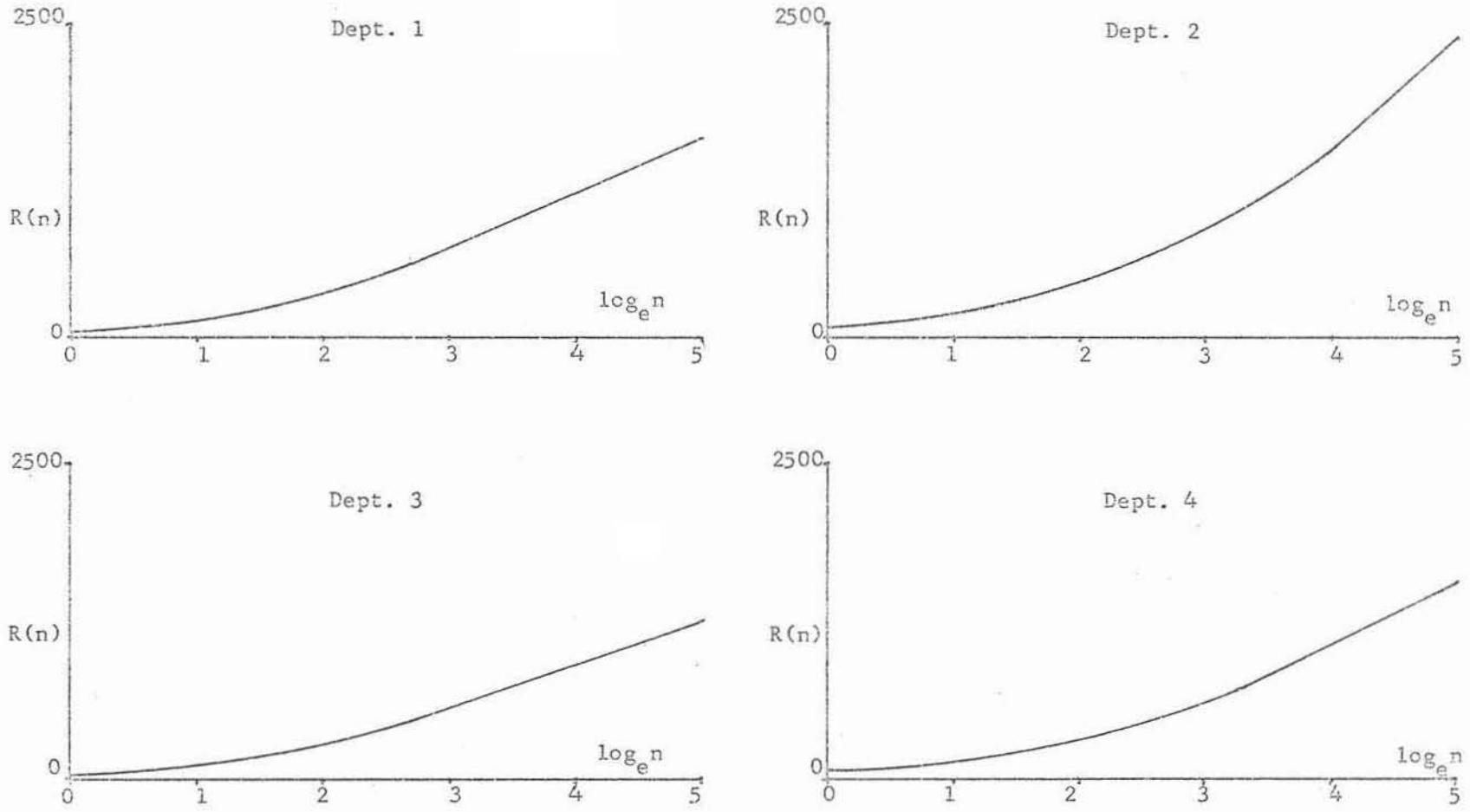

Fig. 5.2

Productivity Distributions

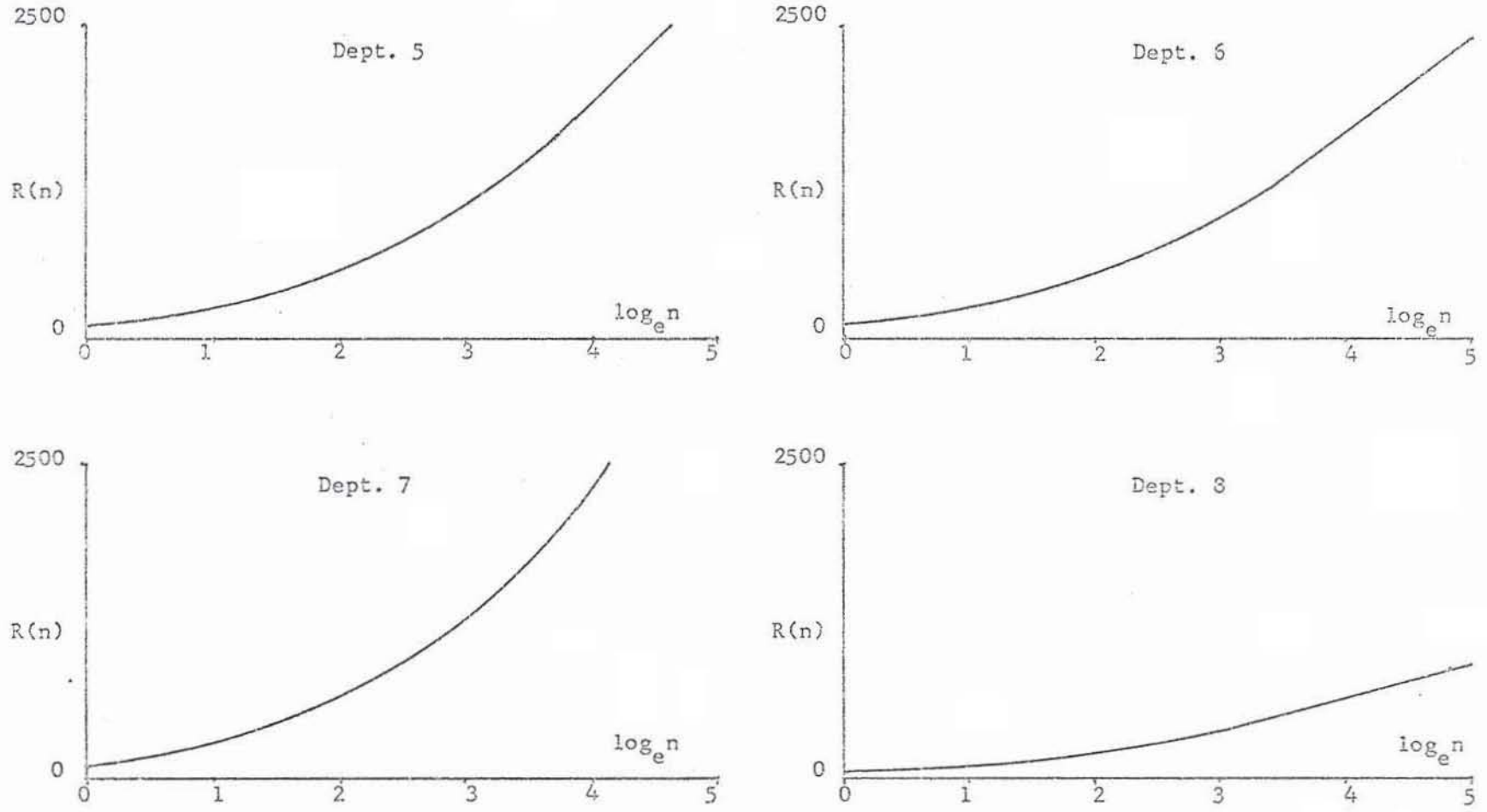

Fig. 5.2 (Continued)

Here, an entry in the (i,j) cell ($i < j$) is the number of journals of common interest to the ith and jth departments.

Summing the first column of Table 5.1 shows the total number of non zero values n_{ij}, to be 1600. The data of Table 5.2 then indicates that the total number of candidate journals is 1320.

5.4.2 Costs

Unbiased frequency distributions of the costs of scholarly and research journals are not available. The closest approximation found was the distribution over prices of the total set of subscriptions processed for College and University libraries by F. W. Faxon Company (Clasquin, 1974).

Rather than base the sample distribution on Clasquin's figures, it was decided that a more useful approach was to generate cost data from a parametric distribution function, to facilitate variation of the cost distribution profile for testing purposes. Normal variates were generated, with a mean of 40 and a standard deviation of 50, but only values greater than 5 were chosen. The distribution of costs used in the only sample problem solved so far, is illustrated in Figure 5.3.

It should be realized that a likely feature of real data sets has been ignored in the present generation of cost data. This feature is the possible existence of correlation between a journal's cost and its productivity rank. No information on the existence of such correlation has been noted, but it would certainly seem a likely possibility.

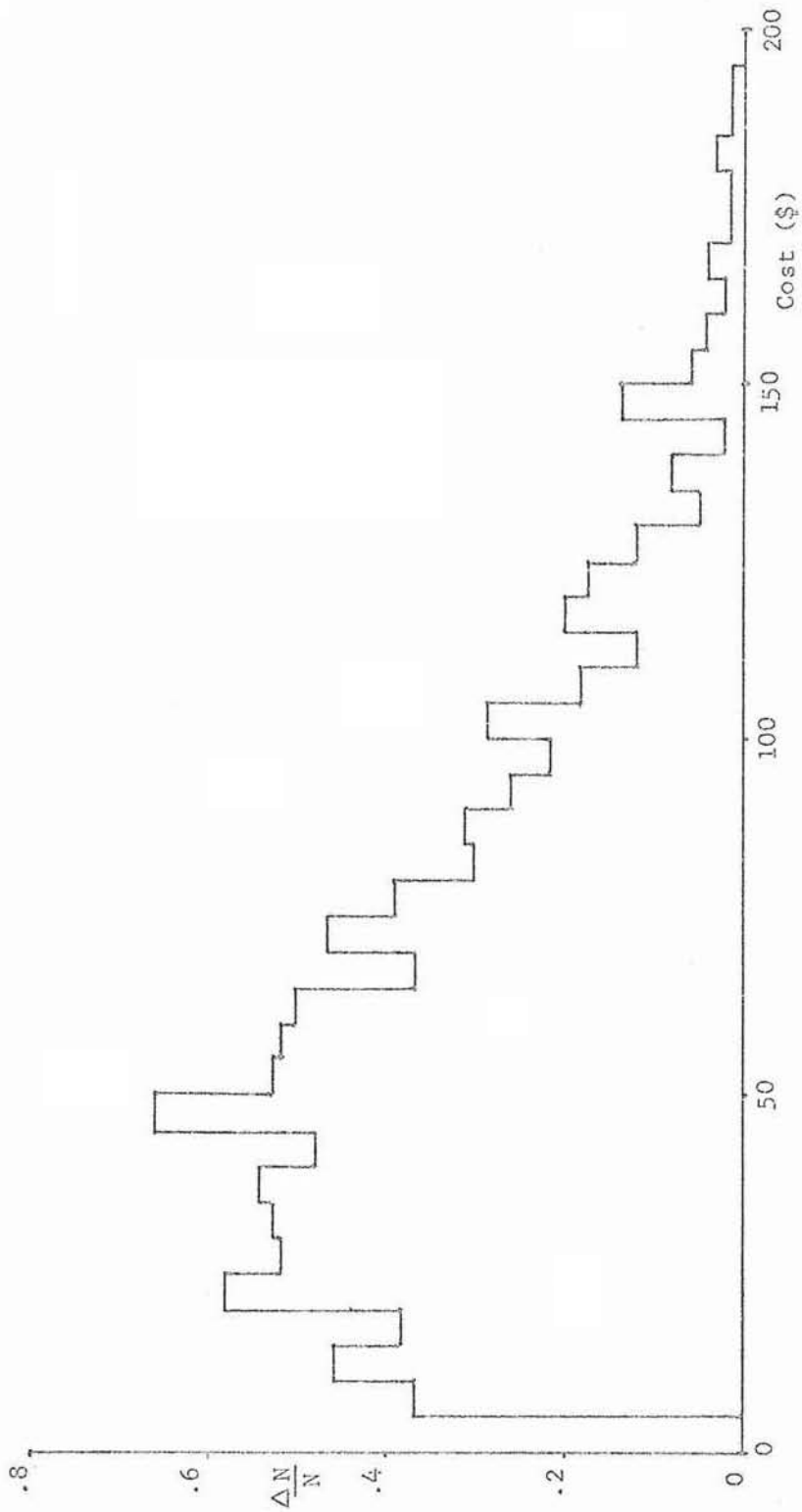

Fig. 5.3

Distribution of Journal Costs

The total cost of the entire set of candidate journals was found to be $79,994, while the greatest possible return was 12,649 articles.

5.5 Results

5.5.1 The Simple Knapsack Problem. Algorithm KNAP1

In agreement with the finding of Ingargiola and Korsh (1973), the reduced problems yielded by the first stage of the algorithm are generally not solved very easily, even though the number of variables is usually not large (e.g., 100). Superficial examination of the variables of some of these reduced problems seems to indicate that this could be because there is very little variation in benefit/cost among them. That is, they are difficult to choose from.

On the other hand, it has been found that in all cases where the algorithm is allowed to continue to optimality, the second stage is found to yield only a very small improvement in the solution, and quite often, no improvement at all. Examples are shown in Table 5.3, where Z^* represents the result of stage 1, and $Z^u_{opt.}$ is the final optimal solution.

TABLE 5.3

B	Z^*	$Z^u_{opt.}$
10,000	5893	5894
22,000	8538	8538

For this reason, the following termination procedure was included in the code.

Normally, a branch is abandoned if its continuous solution is not greater than the incumbent value Z^*. After 500 branches have been examined, in this way, the continuous solution of each ensuing branch is compared with $(Z^* + 1)$, rather than with Z^*. That is, the algorithm demands a possible improvement of at least 2 in the value of the solution before it will persist with the branch.

In practice, this has always forced termination almost immediately after the 500th branch. The final result obtained, $Z^u_{opt.}$, is in fact no worse than one less than the true optimal value - a negligible difference in view of the large values involved.

Two main runs of the algorithm KNAP1 were performed:

(a) A series of solutions was obtained for 21 values of the budget ranging from \$10,000 to \$45,000. This required approximately 90 seconds of CPU time, including input and output. Note that the variables were already ordered according to benefit/cost before being input.

The results are shown in detail in Table 5.4, and graphically in Figure 5.4.

Table 5.4 includes the following information:

Z^* : the best solution obtained in stage 1;

BRANCHES : the number of branches examined;

FRAC ; the ratio $\frac{Z^u_{opt.}}{Z(Max)}$ where Z(Max) is the largest possible return (12649);

$N_{Red.}$: the number of variables in the reduced problem;

Z^o : the value of $\sum_{j \in I_1} c_j$. That is, the total value of

those variables which were pegged to unity in stage 1. Note that these variables might give an objective method of defining a 'core' list of journals - a recurring theme in the literature of journal selection (see, e.g., Scales, 1976).

$B_{Red.}$: the budget available for the reduced problem.

(b) A similar series of solutions was obtained for 20 values of the budget, ranging from 19000 to 20900, in increments of 100. Total CPU time for this run was 102 seconds.

This time, additional output was obtained in the form of the eight values of the return to each department,

$$R_i = \sum_{j \in Q_p} n_{ij} X_j \; ; \quad i = 1, \ldots, M.$$

These returns were expressed in terms of the proportions they represent of the maximum possible returns to each department:

$$FRAC_i = \frac{R_i}{\sum_{j \in Q_p} n_{ij}} \; ; \quad i = 1, \ldots, M.$$

These results are presented in Tables 5.5 and 5.6. Table 5.5 provides the same data as in Table 5.4, while Table 5.6 shows the values of $FRAC_i$, $i = 1, \ldots, M$. A graphical representation is given in Figure 5.5.

It is interesting to note that these solutions for the problem unconstrained by equity considerations display a range of variation in departmental satisfaction ranging from about .5 to about .7.

TABLE 5.4

B	BRANCHES	Z^*	$Z^u_{opt.}$	FRAC	$N_{Red.}$	Z^o	$B_{Red.}$
10000	88	5893	5894	.4660	86	5616	922
12000	502	6441	6441	.5092	93	6027	1515
14000	501	6938	6939	.5486	192	6163	2915
16000	501	7385	7385	.5838	98	7073	1401
18000	501	7797	7797	.6164	49	7603	957
20000	0	8183	8183	.6469	1	8178	27
22000	442	8538	8538	.6750	27	8434	599
24000	501	8868	8868	.7011	135	8461	2354
26000	1	9175	9175	.7253	2	9162	89
27000	502	9315	9316	.7365	198	8757	3548
28000	502	9452	9452	.7473	28	9371	598
29000	501	9583	9583	.7576	48	9409	1318
31000	501	9830	9831	.7772	265	9057	5517
33000	501	10064	10064	.7956	60	9894	1471
35000	501	10281	10281	.8128	42	10157	1156
37000	501	10481	10481	.8286	102	10229	2462
39000	501	10666	10666	.8432	62	10527	1521
41000	502	10839	10839	.8569	20	10791	569
43000	501	10998	10998	.8695	34	10935	812
4500	98	11149	11149	.8814	13	11131	246

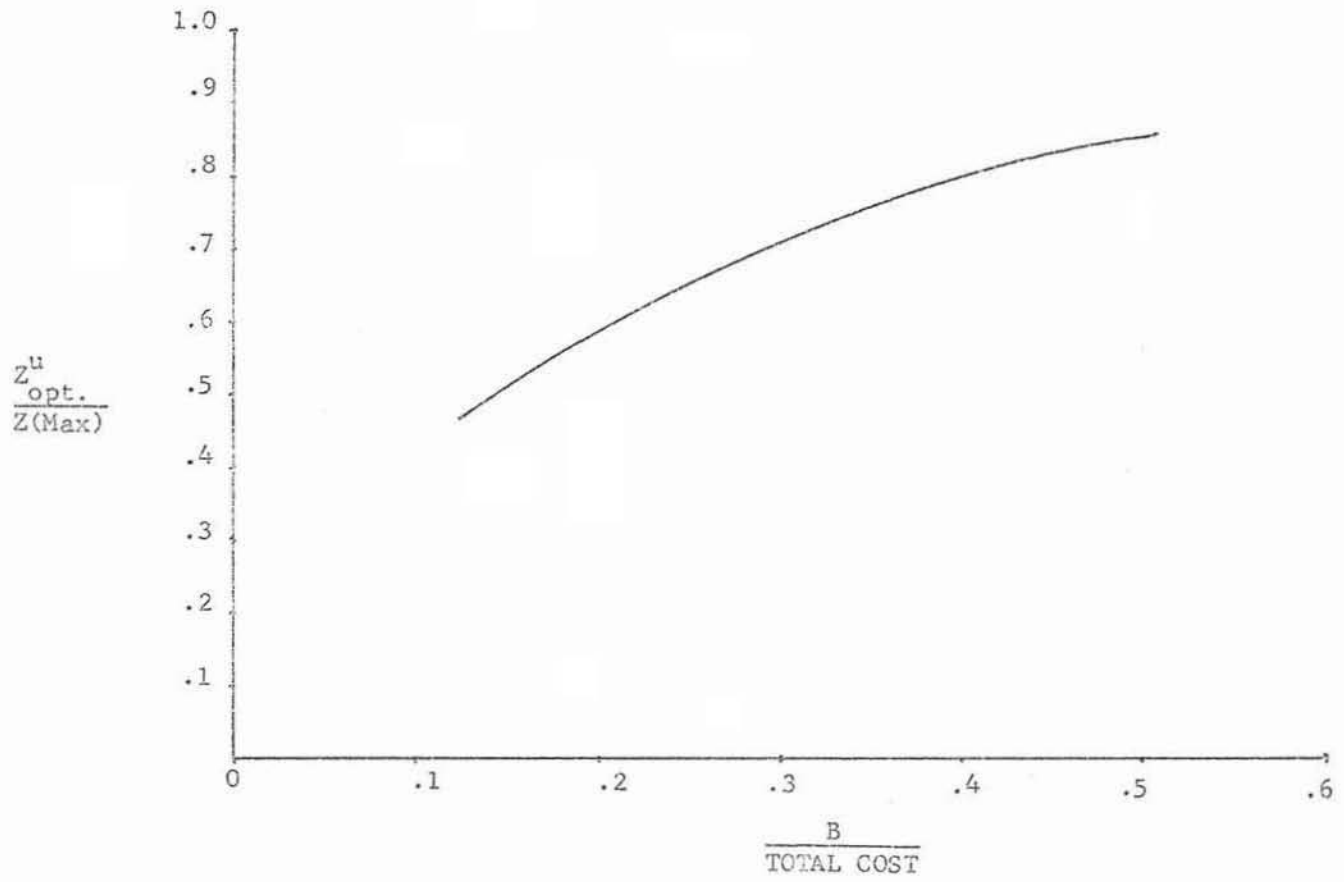

Fig. 5.4

Benefit-Cost Curve: No Equity Constraints

TABLE 5.5

B	BRANCHES	Z^*	$Z^u_{opt.}$	FRAC	$N_{Red.}$	Z^o	$B_{Red.}$
19000	501	7994	7994	.6320	71	7699	1466
19100	502	8013	8013	.6335	103	7611	1972
19200	501	8031	8032	.6350	110	7620	2021
19300	501	8051	8051	.6365	117	7615	2143
19400	501	8070	8070	.6380	114	7660	2019
19500	501	8089	8089	.6395	108	7676	2047
19600	501	8108	8108	.6410	77	7831	1401
19700	501	8127	8127	.6425	38	8022	551
19800	501	8145	8145	.6439	132	7680	2317
19900	502	8164	8164	.6454	54	7995	884
20000	0	8183	8183	.6469	1	8178	27
20100	501	8201	8201	.6484	50	8017	976
20200	502	8219	8219	.6498	92	7942	1431
20300	501	8237	8237	.6512	126	7810	2167
*20400			8256				
20500	501	8273	8273	.6540	141	7821	2301
20600	501	8291	8291	.6555	130	7867	2176
20700	501	8309	8309	.6569	118	7920	2024
20800	501	8327	8327	.6583	103	7949	1996
20900	501	8344	8345	.6597	208	7600	3742

(* First continuous solution also integer.)

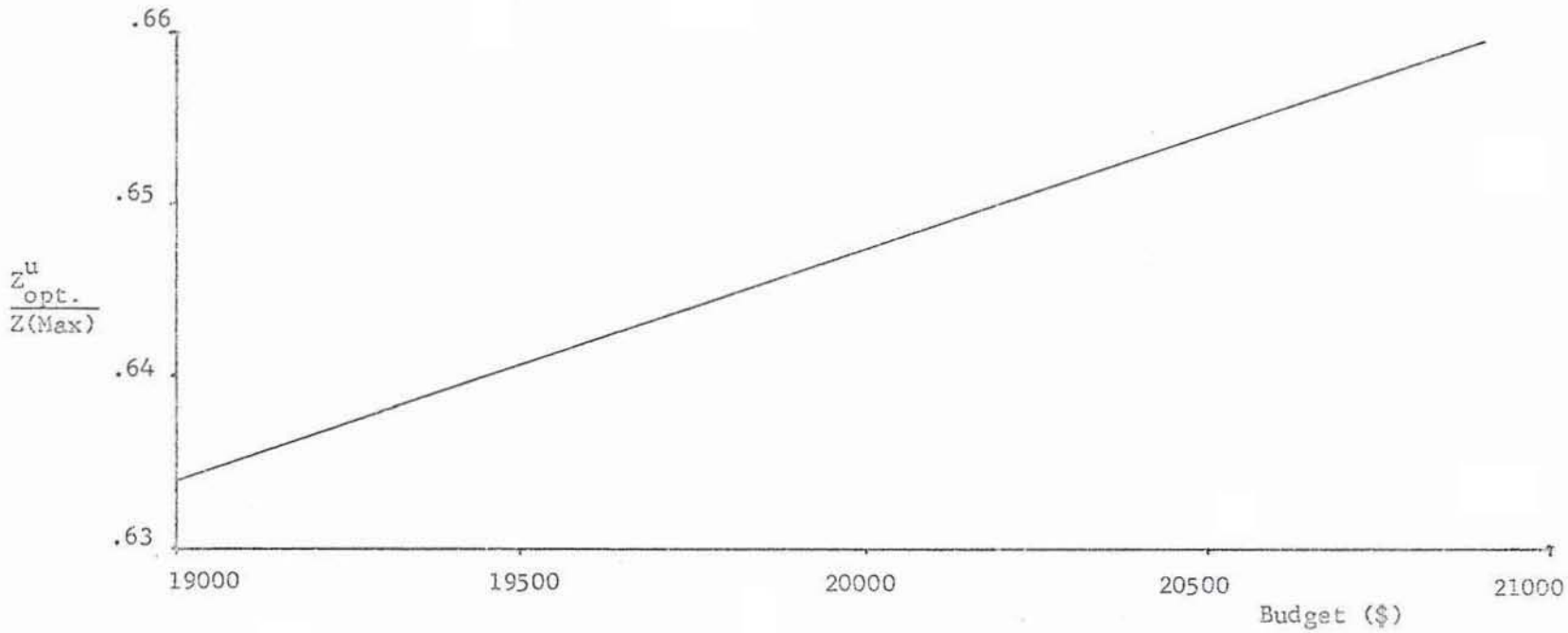

Fig. 5.5

Detail from Benefit-Cost Curve: No Equity Constraints

TABLE 5.6

Dept.	$FRAC_i$(19500)	$FRAC_i$(20000)	$FRAC_i$(20500)
1	.7076	.7215	.7215
2	.6061	.6061	.6137
3	.6263	.6497	.6589
4	.6778	.6851	.7132
5	.6810	.6810	.6810
6	.6169	.6196	.6232
7	.6418	.6516	.6588
8	.5093	.5296	.5465

5.5.2 The Problem Including Equity Constraints. Algorithm JNLSEL.

Again, attention is focussed on expenditure in the region of 20,000 dollars. From Table 5.4, we note that the equity factor q certainly cannot exceed .6469, and in fact should be rather less because some trade-off should occur between equity and total return.

The algorithm JNLSEL was executed once, with a fixed budget of $20,000, and produced output for 20 values of q ranging from .635 to .654, in increments of .001. Total CPU time required was 25 seconds, including input and output. Again, the variables were already sorted in order of benefit/cost, before input.

The results are presented in detail in Tables 5.7 and 5.8.

Here, Z^* represents the best solution value found; OPT is the ratio $Z^*/Z^u_{opt.}$; EXP is the total expenditure required for that solution; and $FRAC_i$, $i = 1, \ldots, M$, have the same meaning as in Table 5.6.

TABLE 5.7

q	z^*	EXP	OPT
.635	8167	20000	.9980
.636	8166	19999	.9979
.637	8164	19998	.9977
.638	8164	19993	.9977
.639	8161	19993	.9973
.640	8157	19995	.9968
.641	8157	19997	.9968
.642	8151	19994	.9961
.643	8143	19992	.9951
.644	8151	20036	
.645	8164	20115	
.646	8173	20168	
.647	8187	20226	
.648	8199	20308	
.649	8211	20357	
.650	8221	20421	
.651	8237	20500	
.652	8249	20561	
.653	8256	20619	
.654	8276	20696	

TABLE 5.8

q : Dept.	.635 $FRAC_i$	.636 $FRAC_i$	.637 $FRAC_i$	.638 $FRAC_i$	.639 $FRAC_i$	.640 $FRAC_i$	.641 $FRAC_i$	.642 $FRAC_i$	.643 $FRAC_i$
1	.6724	.6724	.6628	.6724	.6403	.6403	.6403	.6414	.6489
2	.6356	.6356	.6368	.6381	.6387	.6412	.6412	.6418	.6425
3	.6354	.6385	.6487	.6385	.6385	.6487	.6487	.6456	.6426
4	.6840	.6778	.6674	.6705	.6861	.6747	.6642	.6674	.6475
5	.6570	.6570	.6536	.6512	.6536	.6458	.6497	.6438	.6428
6	.6355	.6364	.6369	.6378	.6387	.6396	.6405	.6424	.6428
7	.6367	.6367	.6397	.6385	.6397	.6418	.6418	.6418	.6430
8	.6345	.6345	.6362	.6413	.6413	.6413	.6413	.6413	.6464

Note that the algorithm finds a feasible solution for a value of q as high as .643, which is not very different from the value of .6469 above which certainly no feasible solution exists. Hence the weakness of the algorithm in that it can fail to find a feasible solution when one does in fact exist, does not appear to be a matter for great concern.

The trade-off which occurs between equity and total return is demonstrated in Table 5.7, as the total return slowly decreases as q increases, and the range of variation of the $FRAC_i$ decreases.

For $q > .643$, the algorithm still provides useful information in the form of low-cost, almost perfectly equitable solutions requiring expenditure in excess of 20,000 dollars.

An important feature of the results is the degree to which the

constrained solutions approach the optimal solution of the unconstrained problem. This is indicated by the high values of OPT, which indicate that the algorithm is very effective in finding highly equitable solutions at a very low trade-off in terms of total return. Of course it should be realized that this effectiveness could result from the special structure of the problem being considered, especially since the solutions obtained for the unconstrained problem could not be regarded as being extremely inequitable (Table 5.6). Nevertheless, results indicate that JNLSEL is an efficient, inexpensive tool for solution of the journal selection problem.

CHAPTER 6

CONCLUDING REMARKS

Progress has been made towards the development of an approach to the problem of journal selection in an academic library. It has been demonstrated that this approach can result in a practical formulation of the problem, for which data requirements should not be impossibly severe. Further, it has been seen that effective solutions to the problem, thus formulated, can be obtained with relative ease.

The journal selection problem has been treated in the context of the total library system, so that expenditure on journals was considered in relation to the contribution it makes to the overall performance measure for the library. Solutions to the narrow problem of selection should ideally provide input for the more strategic analyses required in a systems approach to library management. However it has been possible to give only a qualitive indication, based on a highly simplified model, of the manner in which the potential of the library's holdings can be realized in terms of document exposure. Quantitative models of this process, and of other processes in the library system, will be required before a strategy for the allocation of resources within the library can be developed.

In the meantime, however, the librarian makes decisions regarding the allocation of the budget, on the basis of his experience and expectations. It is felt that an approach of the kind presented here can be of value in this more limited context, since it can provide useful information in at least four ways:

Firstly, by giving an indication of the relative levels of potential document exposure that can be acquired for any given investment. This would provide a framework against which the serials acquisition budget could be decided.

Secondly, by listing feasible sets of titles which, if acquired, would constitute efficient subscription policies.

Thirdly, by giving quantitative meaning to the concept of equity in the provision of journal literature. Support of this nature could be very valuable in discussions with heads of departments.

Finally, by providing a basis for comparison against which the effectiveness of current policies could be evaluated.

There is no suggestion that output from the solution process should provide an inflexible recipe for subscription. Rather, the results of analysis are intended only to provide quantitative support for the librarian's decision-making.

It is likely that the present approach would be of greatest help to the libraries of small provincial universities. Librarians in such situations can be faced with the problem of providing for just as great a diversity of subject interests as are the libraries of the major universities, but on much smaller budgets. Hence their selection problems are likely to be more critical.

So far as implementation of this approach is concerned, a number of questions remain to be answered.

The major difficulty lies with the collection of data relating to journal productivities. As indicated in the text, there is

considerable hope that this problem will be overcome more easily and cheaply as automated retrieval develops. It is difficult to estimate likely costs, but with narrow retrospective searches currently costing from $15 to $100 each, we could expect perhaps an average of $200 per department for the collection of data in this manner. On the other hand, as SDI services become more widely used, data could be collected, at very low cost to the library, by aggregating the regular SDI output received by individual subscribers. Further expense would be involved in analyzing these data to obtain productivity rankings.

More severe difficulties may be encountered in the definition of subject profiles for use in automated searching. These profiles must be neither too broad nor too narrow - a challenging optimization problem exists. Some university departments might insist on a broad statement of interests, thus generating a fictitiously large potential demand and leading to an unfair allocation of resources. Of course this problem is not new to the university librarian, and is a reminder that the subjective element can never be entirely removed from this sort of decision-making.

While there is no fundamental reason why the present approach would not find application in all subject areas within the university, it is recognized that difficulties are likely to be encountered in dealing with the Humanities literature, because it tends to be not so well structured as that of the scientific disciplines. In fact it should be noted that the Bradford distribution itself is established as a feature of the scientific literature only.

In the present formulation, the parameter q, representing the

proportion of all relevant articles that is actually acquired, is taken to be the same for all departments. However, there may arise circumstances under which the librarian sees fit to weight the departments differently. Such an approach should be quite feasible, using the algorithms developed here.

We have considered equity requirements in relation to university departments, rather than to individuals. Hence it is quite possible that certain individuals could ultimately receive much better treatment than others. An expanded problem which considers the equity of individuals is certainly feasible in principle. However the data requirements could be considerably greater. Also, special situations can occur, such as that of the individual whose requirements can be almost wholly met by the acquisition of one title only, so that in any given solution his equity will be either close to zero, or close to 100%.

Since the evaluation of a journal is based on its past productivity, there seems to be no way in which new titles can be incorporated in the present formulation.

Finally, how often should the exercise be carried out? The implication has been made that the analysis can provide an efficient subscription policy for the ensuing 'decision period', at the end of which it is repeated, and appropriate changes are made. As time passes, there may be relative changes in journal prices and productivities, and changes in users' interests. Thus a subscription policy that was appropriate at one time might not be so suitable a couple of years later. The librarian's perception of how quickly circumstances are changing, and the effort involved in the exercise, should determine the frequency with which the analysis is performed.