Information Processing & Management Vol. 26, No. 3, pp. 413–431, 1990
Printed in Great Britain.

0306-4573/90 $3.00 + .00
Copyright © 1990 Pergamon Press plc

JOURNAL SELECTION MODEL: AN INDIRECT EVALUATION OF SCIENTIFIC JOURNALS

VESNA OLUIĆ-VUKOVIĆ and NEVENKA PRAVDIĆ
Institute for Information Sciences (formerly: Referral Centre of the University of Zagreb), P.O. Box 327, 41001 Zagreb, Croatia, Yugoslavia

(*Received* 23 *April* 1989; *accepted in final form* 26 *July* 1989)

Abstract—A general model for the selection of scientific journals based on ranking of the data sources is presented. The validity of the concept applied is supported by multiple testings of the model for journals in the field of chemistry. Several analyses, including the impact of input values on the formation of the model output, are performed.

1. INTRODUCTION

In a previous paper [1] we described a general methodological framework for the selection of scientific journals. The procedure was developed in search of an objective and effective method which would operate within the limits of the library-network in a defined region. Based on the multiplicity principle, it incorporated a number of journal input lists obtained by different criteria for journal selection, i.e., data sources of different types and different origins were used.

The fact that combining data from various sources is necessary made itself most conspicuous by the findings of a comprehensive critical review of sources and methods for journal rankings. The review indicated that the correlations between different methods are rather low [2], and suggested that strict adherence to a strategy based on only one particular selection procedure would not meet many requirements. The comparison of journal lists selected by five techniques [3] confirmed the knowledge that different techniques produce different results. In developing library collections, it is almost taken for granted that not a single one of the available journal selection methods is fully adequate when used on its own; but, it can be helpful when supplemented by other approaches [4]. A literature survey of numerous recent studies on journal selection which combine at least two methods by applying data from different sources was presented in our first paper. It seems, however, appropriate to point at the extremely valuable annotated bibliography of *cca* 100 items relating to collection evaluation in academic libraries [5].

The need for refining journal selection techniques and further search for satisfactory selection principles which would yield a reliable procedure are still urgent. There have been a few attempts at modelling some of the aspects of the journal selection decision problem [6–8]. A comparison of several mathematical models designed earlier for library collections was performed [9]. As crucial issues in journal selection procedure, Kraft [9] emphasized the evaluation mechanism of journal worth as well as the importance of the proper choice of the criteria for selection. A multiattribute approach was recommended. This suggestion came from an earlier paper by Rush et al. [10] in which a set of weighted factors were taken to quantify the value of a specific journal. Usage, relevance, availability elsewhere, and capital investment were specified as measures of journal worth. For a more systematic journal selection decision, Koenig [11] proposed a combination of four basic components: citation data, conventional journal data sources, utility/cost ratio, and journal ranking techniques. In designing a model to ease the cancellation decisions, Broude [12] suggested a multifaceted approach including even seven factors as the basis for determining the worth of a journal. According to the model of Dhawan et al. [13] the overlap between three different types of data sources (coverage by secondary literature, citation data, and use data) was suggested as an indicator of journal relevance. In a recent paper, He and Pao [14] proposed a journal selection and ranking algorithm based on the combined use of cited and citing journals.

In our approach, the aim is to generate a balanced journal collection for a particu-

lar discipline, by the application of overlapping technique. The first prerequisite is the combination of data:

a. From international sources (coverage by the Institute for Scientific Information, Philadelphia, USA; coverage by abstracting services; library holdings of international agencies, of leading universities, or of regional/national periodicals centers, etc.) and,
b. from regional sources (data base processing: publishing habits of native scientists, use and user analyses; expert opinion, etc.) [1].

In the present paper, the appropriate model is elaborated in order to act as a filtering system. The capability of the proposed model to provide a set of relevant journals for a specific discipline is tested on an example of chemistry journals followed by input/output analyses.

2. MODELLING APPROACH

The overlapping technique for the selection of scientific journals is a common expression denoting procedures dealing with a given combination of journal input data taken from different sources. Although these procedures rely on the principle that there is a direct connection between pertinence and duplication (the journals which appear in more input lists presumably being of higher importance), the final goal—proper selection of journals which would constitute an adequate collection nucleus—is by no means a unified process. Several approaches could be differentiated:

1. The frequency approach, taking as the most relevant only those journals which appear in all of the data sources used,
2. an approach taking as the most relevant those journals which are present in all the data sources, together with journals which appear in some, arbitrarily chosen, combinations of sources,
3. an approach including the ranking of the data sources.

Ranking of the data sources, which we introduce as the basis for selection, is supported by a few quantitative criteria derived from the distribution pattern of all the journals among the data sources used. The stress is actually on all the journals, irrespective of the frequency of their occurrence. The purpose of modelling, described here, is to eliminate any subjectivity in the journal selection processing on the one hand, and to avoid the stupendous task of single journal ranking on the other. The final effect of the procedure is, however, an indirect evaluation of scientific journals. The consequence of such an attitude is that each of the journals in the nucleus is considered to be of equal value.

3. FORMULATION OF THE MODEL

Suppose there were a pool containing x number of journals dispersed among n data sources used (A, B, C, D, ... X). (For a detailed description of the pool formation see the next section, under "Case Study"). The status of each journal can be expressed by its occurrence in a given number of the data sources: There is always a set of journals occurring in only one source, a set appearing in two, three up to n data sources. The intention is not to consider each journal separately, but to take as relevant appropriate journals' groupings, each of which contains a certain number of journals; these groupings (AB, AC ... AX, ABC ... ABX ... ABCX.) are named the combinations (c). In other words, it means that the combination AB represents the set of journals occurring in sources A and B. The total number of c depends on the number of the data sources, $c = 2^n - 1$. Only a limited number of these combinations contain journals which can be defined as the most important for a specific discipline; the notion "the active combinations" (a_c) is ascribed to them.

The major problem in modelling is to provide information about the relevance of any particular combination. To solve this, appropriate indicators which could reflect the relative importance of each of the data sources should be developed. The starting point would then be a closer analysis of the distribution pattern of journals among the data sources: Precisely, this means that the overlapping data should be arranged according to the journals' occurrence in the sources (A–X). With the assumption that the journals of greater importance would be found in input lists of a greater number of data sources used, it seems appropriate that a number of higher overlappings be taken as the reference ones. In order to assess the relative importance of the sources, it is suggested to inspect the relation between the occurrence of a given source in reference—and in all existing—overlappings. In this way the status of each data source can be defined by two quantitative indicators:

I_1—partial ratio in the total number of all existing overlappings,

I_2—partial ratio in the reference overlappings.

To calculate these two indicators it is necessary to determine the portion of journals from each source in each overlapping, i.e., in single (1×), in two- (2×), in three- (3×), up to n-fold ($n\times$) overlappings (Scheme 1). By summing of:

a. the rows, the corresponding numbers of input journals in each of the sources (subtotals S_A, $S_B \ldots S_X$), and
b. the first column, the total number of all overlappings (S_Σ), is obtained (Scheme 1):

$$S_A = A/1\times/ + A/2\times/ + A/3\times/ + \ldots A/n\times/$$

$$S_B = B/1\times/ + B/2\times/ + B/3\times/ + \ldots B/n\times/$$

$$\begin{matrix} S_C \\ \vdots \\ \dfrac{S_X}{S_\Sigma}. \end{matrix}$$

The quotients S_A/S_Σ, $S_B/S_\Sigma \ldots S_X/S_\Sigma$ are assigned to the first indicator $I_{1A}, I_{1B}, \ldots, I_{1X}$.

The second indicator shows the proportion of each data source in reference overlappings which are defined as the upper half of the existing overlappings, e.g., in the case of seven-data sources, with seven-fold overlappings as the maximum, the reference overlappings should be those from four-fold upward (Scheme 2):

$$R_A = A/4\times/ + A/5\times/ + A/6\times/ + A/7\times/$$

$$R_B = B/4\times/ + B/5\times/ + B/6\times/ + B/7\times/$$

$$\begin{matrix} R_C \\ \dfrac{R_X}{R_\Sigma}. \end{matrix}$$

The quotients R_A/R_Σ, $R_B/R_\Sigma \ldots R_X/R_\Sigma$ are designated as $I_{2A}, I_{2B} \ldots I_{2X}$.

The proportion of the two indicators gives the relative weight, w_r, for each of the sources, $w_{rA}, w_{rB} \ldots w_{rX}$:

$$w_r = I_2/I_1. \tag{1}$$

These weight parameters are further applied for ranking of the data sources. First, a division into categories is required; the purpose of categorization is to put similar sources

into the same group. The sources are distributed in an appropriate number of categories according to their relative weights, from $w_{r,\max}$ to $w_{r,\min}$. To get a better selectivity, we propose that the difference between the categories must not be smaller than the average difference of the relative weights among the sources. The number of categories (y) and the span (s) must be in accordance with the formula:

$$s = (w_{r,\max} - w_{r,\min})/y \qquad s > \bar{w}_r \ . \tag{2}$$

For categories which accommodate more than one data source, the mean w_r is calculated. The categories are estimated in terms of their weights so that the proportionalities found between the categories are retained in the ponders allocated to the data sources. The weights are then adjusted so that the cumulation of normalized ponders (p) equals ten:

$$\sum_{n=1}^{x} p_n = p_A + p_B + \ldots p_X = 10.0. \tag{3}$$

Whereas the ponders p_A–p_X denote the values of single combinations in the sample, the values of all other combinations (p_c), in the range from two- to n-fold, can be obtained by summing the corresponding ponders:

$$p_c = \sum_{n=1}^{x} p_n \qquad 0 < \Sigma p_c \leq 10. \tag{4}$$

It appears that the value of any combination depends on its multiplicity level (number of data sources included, i.e., order of combination) as well as on the rank of the sources.

Next is the selection of those combinations which contain the most relevant journals, i.e., the active combinations, a_c. Starting from the logical presumption that the quantity in terms of the number of journals decrases with the increase of multiplicity level, a rather broad span of combinations is envisaged, and the following requirement for the active combinations is stated:

$$\sum p_{ac} = 2/3 \max \sum p_n. \tag{5}$$

This would define an a_c as each combination in the interval ranging from $a_{c,\min}$ to $a_{c,\max}$. In the present model this means that the active combinations are those with the value in the interval:

$$3.3 \leq p_{ac} \leq 10. \tag{6}$$

Multiple testing of such a claim is presented in the following section.

4. TESTING OF THE MODEL

To test the validity of the general model proposed for the selection of scientific journals in a particular discipline, an experimental study was undertaken for the field of chemistry. That chemistry is the field of our choice is supported not only by the availability of data for chemistry journals obtained by other selection methods but, as stated by Rice [3], because the selection decisions are nowhere more difficult than in the field of chemistry.

Rank distribution analysis based on the acquired experimental data was also performed. In addition, data from a number of computer-aided simulations are presented.

4.1 *Case study of chemistry journals*

4.1.1 *Data sources—formation of the pool.* Following the methodology and guidelines presented in the previous paper [1], the data from five sources (A–E), of international ori-

gin and of different types, were combined with the data from two regional sources (F, G). The instructions applied for the preparation of input lists are as follows:

4.1.1.1 *Source A (belongs to the Type I).* From *Science Citation Index*, Source Publications—journals arranged by Subject Category [15] a selection of categories relevant for chemistry was made (Biochemistry and Molecular Biology; Chemistry; Chemistry, analytical; Chemistry, inorganic and nuclear; Chemistry, miscellaneous; Chemistry, organic; Chemistry, physical; Crystallography; Electrochemistry; Physics, atomic, molecular and chemical; Polymer science; and Spectroscopy). Full coverage of journals listed under these categories was taken. In this way the first 422 journals were identified and introduced into the pool; each journal was accompanied by the label A.

4.1.1.2 *Source B (Type II).* Served the *Chemical Abstracts.* Since it is a product of a comprehensive abstracting service which processes more than 14,000 primary journals relevant to chemistry and chemical engineering, only an arbitrarily taken fraction of the total coverage seemed to be adequate for this investigation. The size of the fraction to be taken was directed by the size of the input list from A. Therefore, from the list of 1000 Most Frequently Cited Journals in Rank Order [16], the journals ranked from 1 to 426 were labeled B and added into the pool.

4.1.1.3 *Source C (Type II).* Full list of journals covered by the selective abstracts journal *Chemischer Informationsdienst* [17], consisting of 189 items, was taken. Each journal was marked by C before entering the pool.

4.1.1.4 *Source D (Type III).* From the library holdings of a leading European university [18], journals appearing under the UDC (Universal Decimal Classification) from 54 to 548 were selected. The obtained input list contained 938 journal titles. In our examination they carry the label D.

4.1.1.5 *Source E (Type III).* From the library holdings of a foreign national periodicals center [19] titles listed under Chemistry, Chemical industry, Food science and technology, Pharmaceutical industry, and Polymers, were taken. The list contained 793 journals which were introduced with the label E.

The data used as the sources of regional origin (F and G) were collected in the course of several empirical studies performed earlier. In both cases they contributed to the better understanding of the selective needs for chemistry journals on regional basis, what was in fact the main motive for their incorporation into the model construction [1].

4.1.1.6 *Source F.* Here belong 253 journals identified as carriers of relevant information for the users of selective dissemination of information (SDI) from the CASEARCH data base [20]. These journals were labeled F and added to the pool.

4.1.1.7 *Source G.* This source of data reflects the publishing habits of the native scientists who are predominantly oriented towards publishing in chemical journals printed abroad. This actually means that chemists from Croatia do need the journals to which they actively contribute. Data were collected:

1. by examining overall output on a large sample of chemists [21], and,
2. by checking the diffusion of dissertation contents into the open literature on a sample of scientists with Ph.D. in chemistry defended at the University of Zagreb [22,23].

Lists of foreign journals identified in two investigations were merged to give 244 titles; they were marked by G and included into the pool.

The pool now contained 1,833 items, denoted by P. The number of theoretical combinations for $n = 7$ gives $c = 127$. The survey is given in Table 1; empty positions are included as well.

4.1.2 *Identification of the active combinations.* Numerical data cumulated in Table 1 were then organized so as to demonstrate the distribution of journals in each of the seven sources, in seven possible overlappings, in the way shown in Schemes 1 and 2. Indicators I-1 and I-2 were calculated and the corresponding relative weights, *wr*, were obtained. All these data are summarized in Table 2.

Table 1. Survey of the combinations (c = 127) for the case study of chemistry journals formed in the pool of seven data sources (A–G), according to their multiplicity level (single and two- to seven-fold overlappings). Data denote the number of journals in a given combination. Combinations marked by (*) are the active combinations (a_c)

		Overlappings					
	Single	Two-fold	Three-fold	Four-fold*	Five-fold*	Six-fold*	Seven-fold*
No.		/8–28/	/29–63/	/64–99/	/100–119/		
/1/	(A) 24						
/2/	(B) 155						
/3/	(C) 12						
/4/	(D) 492						
/5/	(E) 416						
/6/	(F) 59						
/7/	(G) 65						
		(AB) 1	(ABC) 0	(ABCD) 0	(ABCDE) 4		
		(AC)* 1	(ABD)* 21	(ABCE) 0	(ABCDF) 0		
		(AD) 101	(ABE)* 1	(ABCF) 0	(ABCDG) 1		
		(AE) 8	(ABF)* 1	(ABCG) 0	(ABCEF) 2		
		(AF)* 1	(ABG)* 1	(ABDE) 10	(ABCEG) 0		
		(AG) 0	(ACD)* 9	(ABDF) 4	(ABCFG) 0		
		(BC)* 1	(ACE)* 1	(ABDG) 5	(ABDEF) 11		
		(BD) 5	(ACF) 0	(ABEF) 1	(ABDEG) 7		
		(BE) 18	(ACG) 0	(ABEG) 0	(ABDFG) 2		
		(BF) 25	(ADE) 43	(ABFG) 0	(ABEFG) 1		
		(BG) 18	(ADF)* 5	(ACDE) 24	(ACDEF) 4		
		(CD) 4	(ADG)* 8	(ACDF) 1	(ACDEG) 13		
		(CE) 19	(AEF) 0	(ACDG) 0	(ACDFG) 0		
		(CF) 0	(AEG)* 1	(ACEF) 0	(ACEFG) 0		
		(CG) 0	(AFG) 0	(ACEG) 2	(ADEFG) 6		
		(DE) 35	(BCD) 0	(ACFG) 0	(BCDEF) 1		
		(DF) 2	(BCE)* 5	(ADEF) 6	(BCDEG) 0		
		(DG) 4	(BCF) 0	(ADEG) 13	(BCDFG) 0		
		(EF) 7	(BCG)* 3	(ADFG) 0	(BCEFG) 8		
		(EG) 2	(BDE) 2	(AEFG) 0	(BDEFG) 1		
		(FG)* 3	(BDF)* 1	(BCDE) 1	(CDEFG) 0		
			(BDG) 0	(BCDF) 0			
			(BEF)* 14	(BCDG) 0			
			(BEG)* 4	(BCEF) 4			
			(BFG)* 3	(BCEG) 1			
			(CDE)* 6	(BCFG) 0			
			(CDF)* 1	(BDEF) 4			
			(CDG) 0	(BDEG) 0			
			(CEF)* 1	(BDFG) 0			
			(CEG)* 3	(BEFG) 2			
			(CFG) 0	(CDEF) 0			
			(DEF) 2	(CDEG) 0			
			(DEG) 1	(CDFG) 0			
			(DFG) 0	(CEFG) 1			
			(EFG) 0	(DEFG) 0			
/120/						(ABCDEF) 13	
/121/						(ABCDEG) 9	
/122/						(ABCDFG) 1	
/123/						(ABCEFG) 0	
/124/						(ABDEFG) 22	
/125/						(ACDEFG) 1	
/126/						(BCDEFG) 0	
/127/							(ABCDEFG) 32
	1.223	255	137	79	61	46	32

As proposed, the data sources were, according to their relative weights (eqn. 2), placed into four categories: in I, the source C; in II, A, F, and G; in III, B; and in IV, D and E (Table 3). For categories II and IV mean wr were calculated. So that the proportionality found between the categories could be retained in the ponders to be allocated to the data sources, the estimation of the categories in terms of their weights was carried out in two steps:

Table 2. Distribution of chemistry journals in the pool of seven data sources (A–G) indicating partial ratios in all sample combinations (I–1), in reference combinations (I–2), and relative weights (wr) for each of the sources

	← S →	Overlappings										
								← R →				
Sources	Single 1×	2×	3×	4×	5×	6×	7×	S	R	$S/S_\Sigma = I_1$ (%)	$R/R_\Sigma = I_2$ (%)	$I_2/I_1 = w_r$
(A)	24	112	91	66	51	46	32	422	195	12.92	17.40	1.347
(B)	155	68	56	32	38	45	32	426	147	13.05	13.11	1.005
(C)	12	25	29	34	33	24	32	189	123	5.79	10.97	1.895
(D)	492	151	99	68	50	46	32	938	196	28.73	17.48	0.608
(E)	416	89	84	69	58	45	32	793	204	24.29	18.20	0.749
(F)	59	38	28	23	36	37	32	253	128	7.75	11.42	1.474
(G)	65	27	24	24	39	33	32	244	128	7.47	11.42	1.529
Total	1.223							3.265	1.121	100.00	100.00	

Table 3. Adjusted ponders (p) for seven data sources (A–G) used in the case study of chemistry journals

Category	Data sources	Mean w_r	Ponder $w_r/0.678$	Adjusted ponder (p)
I	(C)		2.79	2.2
II	(A), (F), (G)	1.450	2.14	1.7
III	(B)		1.48	1.2
IV	(D), (E)	0.678	1.00	0.75

1. mean value for wr of the sources in the lowest category was taken as ponder 1.00, and it was allocated to the sources D and E. In this relation, ponders for other five sources were calculated;
2. an adjustment was done to fulfill the requirement stated in eqn. (3).

The adjusted ponders, p, given in the last column of Table 3, are: for the data sources A, 1.7; B, 1.2; C, 2.2; D, 0.75; E, 0.75; F, 1.7; and G, 1.7 ($\Sigma p = 10.0$).

The values of the existing combinations, pc, were calculated by summing the ponders of the corresponding sources. Finally the ac were identified according to the eqn. (6), i.e., these were all the combinations with $pc > 3.3$.

In this case, for the journals in chemistry, all the combinations from the reference overlappings fall into the range of the ac. In addition, in the group of the three-fold and even among the two-fold overlappings, there are some of the combinations that satisfy the required condition. These combinations are marked in Table 1.

4.1.3 *Compilation of the nucleus journals list.* The master list of all foreign journal titles occurring in any of the active combinations was compiled. The list of nucleus journals thus obtained contained altogether 313 titles:

- 32 of them originating from seven-fold overlappings, i.e., the journals present in the input lists of all the sources used,
- 46, 61, and 79 journals from the groups of six-, five-, and four-fold combinations, respectively,
- 89 journals selected from the three-fold combinations,
- 6 journals from highly ranked two-fold overlappings belonging to the active combinations.

Once the nucleus journals are selected, the list is organized alphabetically, and each of the journals in the nucleus is considered to be of equal value.

At this point the journals published in Yugoslavia, which fulfill the selection criteria, should be introduced. There are three such journals that according to their occurrence in

Table 4. List of the nucleus journals for chemistry (a fragment) selected by means of the proposed model, using input lists from seven data sources (A-G)

No.	Journal title	(A)	(B)	(C)	(D)	(E)	(F)	(G)
1.	*Accounts of Chemical Research*	+		+	+	+		
2.	*ACS Symposium Series*	+	+		+		+	
3.	*Acta Chemica Scandinavica*, Ser. A	+		+	+	+		
4.	*Acta Chemica Scandinavica*, Ser. B	+	+	+	+	+	+	
5.	*Acta Chimica Hungarica* (formerly *Acta Chimica Academiae Scientiarum Hungaricae*)	+		+	+	+		
6.	*Acta Crystallographica*, Sect. A	+			+	+		+
7.	*Acta Crystallographica*, Sect. B	+	+		+	+		+
7a.	*Acta Pharmaceutica Jugoslavica*			+				+
8.	*Acta Polymerica*	+			+	+	+	
9.	*Advances in Carbohydrate Chemistry and Biochemistry*	+		+	+	+		
10.	*Advances in Catalysis*	+		+	+	+		
11.	*Advances in Chemistry Series*	+		+	+		+	
12.	*Advances in Heterocyclic Chemistry*	+		+	+	+		
13.	*Advances in Inorganic Chemistry and Radiochemistry*	+		+	+	+		
14.	*Advances in Organometallic Chemistry*	+		+	+	+		
15.	*Advances in Photochemistry*			+	+	+		
16.	*Advances in Physical Organic Chemistry*	+		+	+	+		
17.	*Agricultural and Biological Chemistry*	+	+		+	+	+	
313.	*Zhurnal Vsesoyuznogo Khimicheskogo Obshchestva*	+		+	+	+	+	
	Total number of journals	246	202	157	248	242	158	157

the data sources belong to the active combinations (ACDEG, CEG, and CG). The final number of journals in the list (N) is, thus, 316.

Table 4 contains a fraction of the list of 316 nucleus journals for chemistry (a complete list can be obtained on request) and indicates in which data sources these titles are included.

4.2 *Rank distribution of journals—Graphical determination of the nucleus*

A further step of the model testing relies on the distribution pattern of chemistry journals among the total number of existing combinations in the case study described (Table 1). The rank distribution approach was applied; the combination value (eqn. 4) was taken as the ranking parameter. Before the graphical data analysis was performed, two assumptions were postulated: First, by analogy to Bradford law it was presumed that the nucleus journals would be associated with the exponential part of the curve; second, if the presumption about a_c is valid (eqn. 6), the exponential part of the curve would run into a straight line around combination value, p_c, of 3.3.

In order to express the total number of journals (P) in the pool of n data sources which are distributed among all the existing combinations (theoretically $2^n - 1$) the following formula is given:

$$P = \sum_{Rc=1}^{Rc=2^n-1} P_{Rc} \tag{7}$$

where P_{Rc} denotes the number of journals in a combination of a rank Rc.

Using this expression, the experimental data were calculated; arranged according to the combination rank in the decreasing order, they are presented in Table 5. Plotting the cumulative total of journals, P, against the combination rank, Rc, on a logarithmic scale, gives a J-shaped curve (Fig. 1). Exponential part of the curve, which contains the most important journals, runs into a straight line around $Rc = 37$ (corresponds to p_c 3.4, see Table 5), to end with a large group of journals present in only one of the data sources. It follows that the number of journals in the core is 313. Graphical determination of nucleus

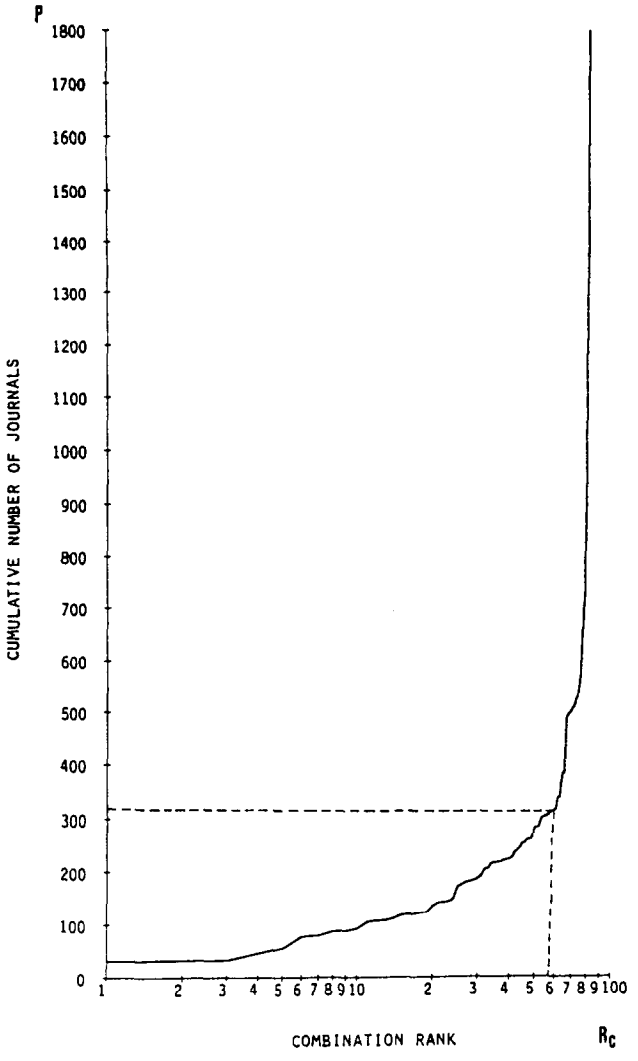


Fig. 1. Distribution of chemistry journals in the case study sample with the data sources A–G, over existing combinations.

journals stands, thus, in agreement with the data obtained in the procedure given in the preceding section, i.e., by the identification of the a_c according to the proposed eqn. 6.

Computer-aided simulations for a number of pools containing input journals from the same data sources ($n = 5$ or 6) were used to perform further testing. The following sets of data sources were considered:

$n = 5$	I	A,B,C,D,E
	II	A,B,C,F,G
	III	A,B,E,F,G
$n = 6$	IV	A,B,C,D,E,F
	V	A,B,C,D,E,G
	VI	A,C,D,E,F,G

As described above, for each set (I–VI), the combination values were calculated, arranged according to combination rank in the decreasing order, associated with the corresponding number of journals, P_{Rc}, and presented graphically in an analogous manner. The number

Table 5. Rank distribution of chemistry journals in the case study sample with seven data sources (A-G), according to the combination rank in decreasing order

Combination rank R_c	Combination code	Combination value, p_c	Number of journals, NR_c	Cumulation P
1	ABCDEFG	10.00	32	32
2	ABCDFG	9.25	1	33
3	ACDEFG	8.80	1	34
4-5	ABCDEF	8.30	13	
	ABCDEG		9	56
6	ABDEFG	7.80	22	78
7-9	ABCEF	7.55	2	
	BCEFG		8	
	ABCDG		1	89
10-11	ACDEF	7.10	4	
	ACDEG		13	106
12-13	ABDFG	7.05	2	
	ABEFG		1	109
14-16	ABCDE	6.60	4	
	ADEFG		6	
	BCDEF		1	120
17-19	ACDF	6.35	1	
	ACEG		2	
	CEFG		1	124
20-22	ABDEF	6.10	11	
	ABDEG		7	
	BDEFG		1	143
23-24	BCEF	5.85	1	
	BCEG		4	148
25	ACDE	5.40	24	172
26-29	ABDF	5.35	4	
	ABDG		5	
	ABEF		1	
	BEFG		2	184
30	BCG	5.10	3	187
31-33	ADEF	4.90	6	
	ADEG		13	
	BCDE		1	207
34-38	ACD	4.65	9	
	ACE		1	
	CDF		1	
	CEF		1	
	CEG		3	222
39-41	ABF	4.60	1	
	ABG		1	
	BFG		3	227
42-43	ABDE	4.40	10	
	BDEF		4	241
44-47	ADF	4.15	5	
	ADG		8	
	AEG		1	
	BCE		5	260
48	AC	3.90	1	261
49	CDE	3.70	6	267
50-54	ABD	3.65	21	
	ABE		1	
	BDF		1	
	BEF		14	
	BEG		4	308
55-57	AF	3.40	1	
	BC		1	
	FG		3	313
58-60	ADE	3.2	43	
	DEF		2	
	DEG		1	359

Table 5. Continued

Combination rank R_c	Combination code	Combination value, p_c	Number of journals, NR_c	Cumulation P
61–62	CD	2.95	4	
	CE		19	382
63–65	AB	2.90	1	
	BF		25	
	BG		18	426
66	BDE	2.70	2	428
67–72	AD	2.45	101	
	AE		8	
	DF		2	
	DG		4	
	EF		7	
	EG		2	552
73	C	2.20	12	564
74–75	BD	1.95	5	
	BE		18	587
76–78	A	1.70	24	
	F		59	
	G		65	735
79	DE	1.50	35	770
80	B	1.20	155	925
81–82	D	0.75	492	
	E		416	1833

of nucleus journals (N) in each of the sets was determined graphically and by using the eqn. 6; these data being incorporated into the Fig. 2.

The data generated for all the testing sets correspond to those established for the case-study sample and, in addition, show that:

a. Distribution patterns of journals expressed as J-shaped curves are typical,
b. the nucleus journals are associated with the exponential part of the curves,
c. the curves run into a straight line around $p_c = 3.3$ (actually in the interval from 3.1 to 3.7),
d. there is a strong correspondence between graphical determination of the nucleus journals and their determination on the basis of the proposed model.

All these facts support the model elaborated in our study confirming the validity of the concept of the active combinations, in general, and the correctness of the span suggested in the eqn. 6, in particular.

Before proceeding, let us turn our attention to the specific features of this distribution. It appears that the concentration-dispersal phenomenon is exhibited here in a special, reversed (inversed) form. Whereas in standard rank distributions (e.g., papers in journals), there are a great number of papers concentrated in a limited number of higher-ranked journals and a small number of papers dispersed over a very large number of less-productive journals, the data in Table 5 clearly show that only a small number of journals are found in higher-ranked combinations and that a great number of journals occupy the low-ranked combinations (including the singles). Due to this reversal, it follows that in order to determine a proper number of items in the core, one has to involve two-thirds of the ranking scale as suggested by eqn. 5.

It should be emphasized that the application of the rank approach in the selection of journals for a specific discipline, as described here, is just another illustration showing the amalgamation of the graphical and verbal formulation of the law of scattering. More examples including appropriate rationalization are presented elsewhere [24]. In passing, let

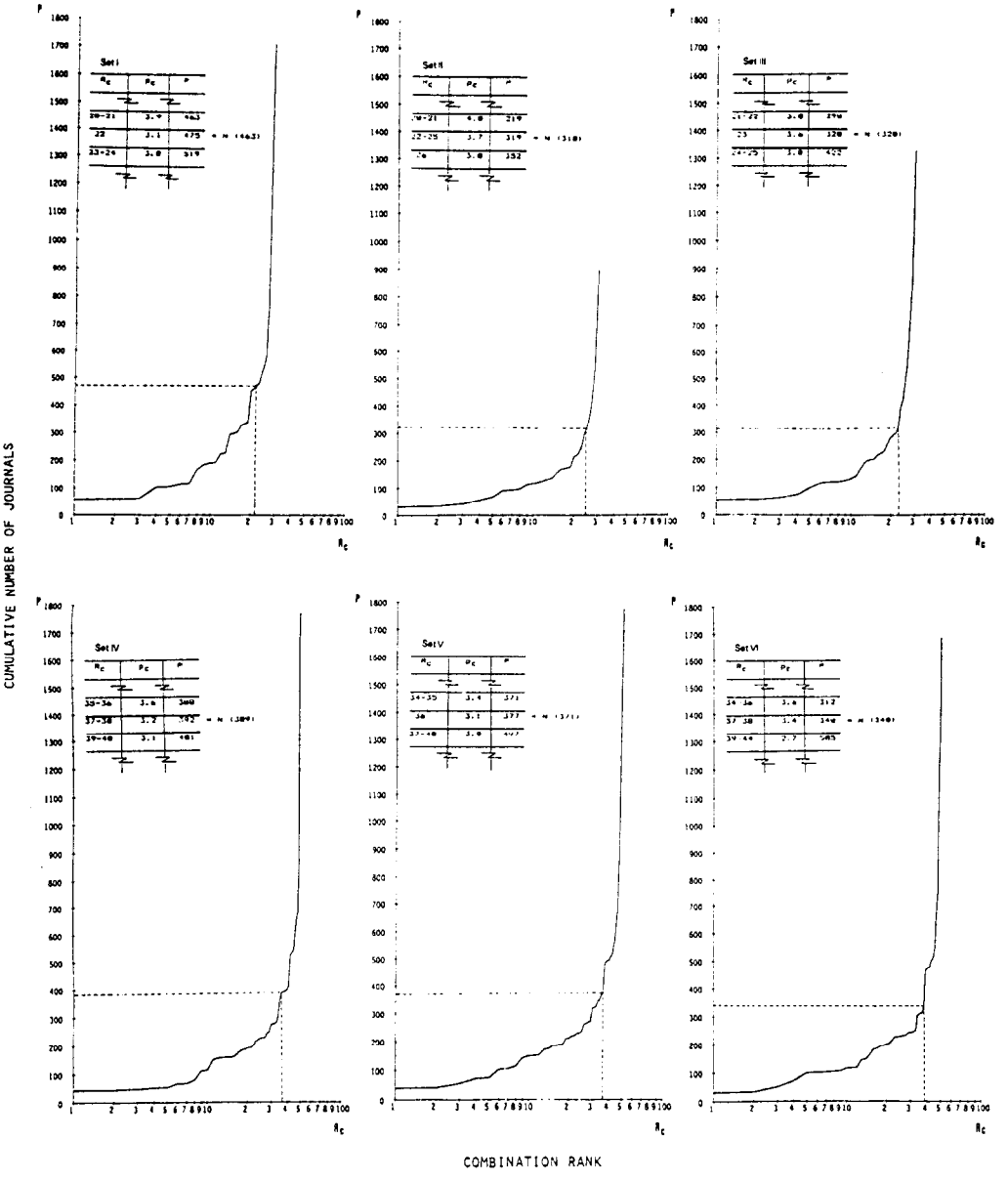


Fig. 2. Distribution of journals in simulated sets I–VI. For determination of the number of nucleus journals (N), graphical presentation should be combined with tabulated data (inserts) containing Rc (combination rank), pc (combination value), and P (cumulative number of journals) for the appropriate regions. Data for N given in parentheses denote the values calculated according to eqn. 6.

us mention one more, also atypical, application of the rank approach, which proved to be useful in determining the most prolific authors in one specific field [25].

5. INPUT/OUTPUT ANALYSIS

The obtained results allow us to perform a few analyses in an attempt to examine the mechanism underlying the process of journal selection. According to the model, the probability of a journal to be a part of the nucleus is primarily a function of:

1. The frequency of its appearance in a number of data sources and,
2. the rank of data sources.

Apart from this, it seems plausible to presume that there must be some additional factors which might be responsible for the formation of a nucleus journals list. Since there is no ground to decide *a priori* which those factors are, an impact evaluation of input values is performed in order to distinguish the factors which affect the process of journal selection from those which do not. Number and size of data sources were considered as input values.

5.1 *Number of data sources (n)*

In the formulation of guidelines for the choice of the data sources we suggested already at the methodological level that five data sources should be taken as a minimum [1]. Such a suggestion was supported by the necessity of combining sources of various types and different origins. The requirements formulated by the model reinforced the earlier proposition. Since the upper and the lower limits determining the active combinations (eqn. 6) are fixed, it could be expected that in a pool containing only three or four data sources, ponder value of a single source might exceed the $a_{c,\min}$. When this happens, the model's selectivity is blocked, with a too numerous nucleus at the outcome.

For this reason, as the basis for the analysis of a possible effect of n, the following sets were used: for $n = 5$, I-III; for $n = 6$, IV-VI; and for $n = 7$, the case-study sample.

If the number of data sources used possesses an impact on the formation of nucleus journals list, it should be reflected on the quantity of output produced by the model. In order to test such a hypothesis, the effect of n on the number of active combinations and, consequently, on the number of journals constituting the nucleus, N, was considered. Two ratios were also analysed:

a. Total number of active combinations towards the total number of existing combinations,
b. number of active combinations from the lowest overlap (two-fold) in relation to the total number of active combinations.

By applying simple regression method for data analysis, the relations of n towards the number of total combinations, active combinations, the ratios (a) and (b) and finally towards N were obtained:

	c	a_c	Ratio (a)	Ratio (b)	N
n	0.972	0.974	−0.729	−0.960	−0.271

As can be seen, both the total number of existing combinations and the total number of active combinations increase proportionally with the increase of the number of data sources. But, if their ratio, (a), is taken into consideration, a tendency of decreasing occurs, indicating that the number of a_c is affected by n. The main reason for this lies in the fact that the portion of the active combinations from the two-fold overlappings significantly decreases by the increase of the total number of active combinations /ratio (b)/. It seems, however, that there is no indication about a possible impact of these changes on N (given in Fig. 2), which was found to vary: 318–463, 340–389, and 316 in collections with $n = 5$, 6, and 7, respectively.

Although it is evident that there is no correlation (-0.271) between n and N, it appears that there are cases where N is almost constant while n changes from 5 to 7, on the one hand, and cases where n is fixed while N changes, on the other. These findings will be commented on later in this paper.

5.2 *Size effect*

Further question concerning the impact of input values on the formation of nucleus journals list arises from the apparent dependency of the rank and source size.

According to the procedure applied in the construction of the model which indirectly uses the source size as one of the criteria for ranking, smaller sources were ranked higher than sources with larger number of journals. This might lead to the presumption that the contribution of sources to the nucleus is, at least in part, affected by their size.

As the sources used in the case-study sample differ considerably in their size, e.g., difference between C and D being 1: 5, this set seemed to be suitable for examination. Two ratios were considered for each of the seven data sources:

- total number of input journals from a particular source in relation to the total number of journals in the pool, (I),
- number of nucleus journals from each source (see data in Table 4) in relation to N, (II).

The obtained results are presented graphically in Fig. 3: both curves exhibit a similar trend of increase from the higher ranked to the lower ranked data sources (with the exception of the fourth data source, curve II). On this basis, one may conclude that the sources with higher ratio of journals/pool tend to have a higher ratio of journals/nucleus. But if the difference between these two ratios for each of the data sources is taken into consideration, it can be seen that the impact of the size is significantly reduced through the process. The larger is the difference between the ratios I and II, the larger is the contribution of a source to the nucleus. This value might then be used as an indicator of the real contribution of each particular source.

In the case of the source A the difference between the ratio I and II is significantly higher in comparison with other data sources. Next is the source B, while the contribution of the sources C, F, G, and E is more or less similar. The contribution of source D is significantly lower. It should be emphasized that this sequence corresponds to the type order, which was in fact attributed to the data sources of international origin with the aim to classify them according to their quality [1].

The obtained results indicate that there is no direct connection between the size of a source and its contribution to the nucleus.

In summary, it appears that the input/output analyses performed with two simple

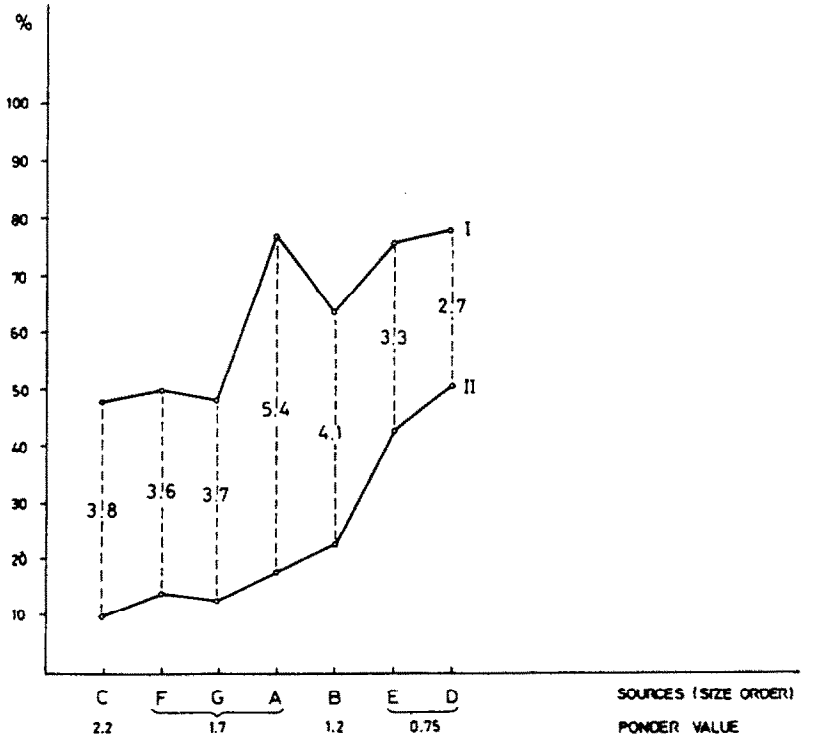


Fig. 3. Input/Output analysis: The size of the data sources A–G used in the case study of chemistry journals in relation to their contribution to the nucleus. I, the ratio of input journals in the pool; II, the ratio in the nucleus.

input values are not sufficient to understand the mechanism operative in the journal selection process. It seems that some other factors originating probably from the complex relations existing among the data sources, as a consequence of their own organization, might be responsible. Therefore, the next step is to examine the relationships between the data sources by studying their interaction and mutual dependency. In order to specify the actual relationship among the sources used in the procedure, the term 'compatibility of sources' is introduced.

5.3 *Effect of compatibility of sources*

Before the analysis of compatibility could be undertaken, a common quantitative measure which enables comparison has to be defined.

If the data sources are considered as aggregating classes ($X_i, X_j, \ldots, X_n$) and journals they contain as the elements of these classes, it seems logical to assume that between any pair of classes there exist a certain number of identical elements (m). This number in relation to the sum of input journals ($\Sigma X_i X_j$) can be used as a characteristic of compatibility. Since the number of identical journals is limited by a number of journals in the smaller data source, for each pair of sources it is necessary to consider the real and the maximal number of identical journals. In this way, degree of compatibility (d) can be obtained, expressed by the equation:

$$d_{XiXj} = \frac{100(\Sigma X_{ij} - m_{\max})}{m_{\max}} \cdot \frac{m_{\text{real}}}{\Sigma X_{ij} - m_{\text{real}}}. \tag{8}$$

Using empirical data summarized in Table 1 for the seven data sources, all theoretically possible pairs of sources, i.e., 21 of them, were examined and corresponding d calculated according to eqn. 8.

The obtained results are presented graphically in the decreasing order (Fig. 4): with

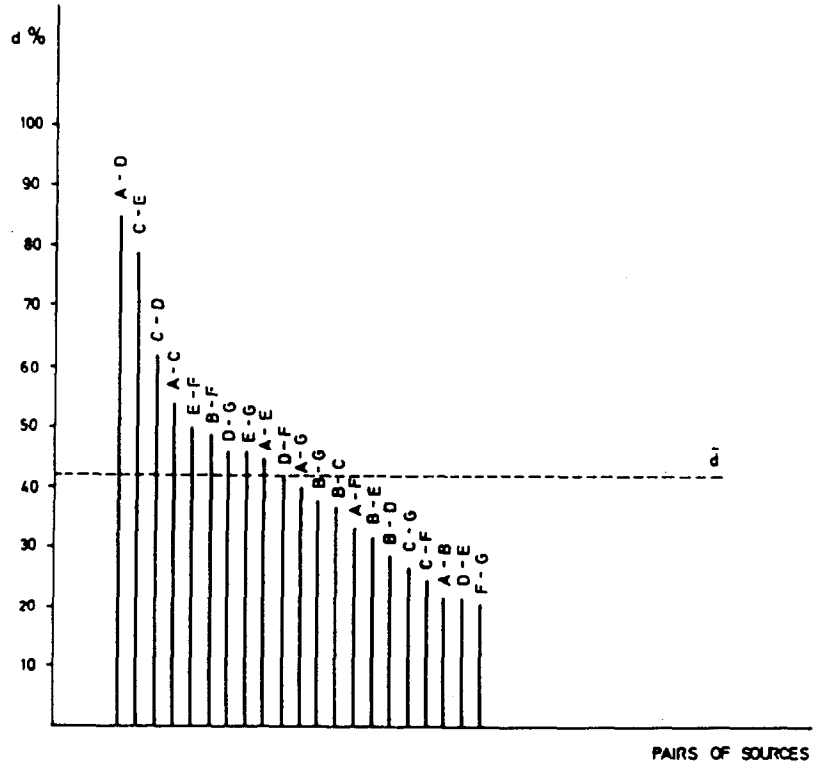


Fig. 4. Input/Output analysis: Degree of compatibility (d) according to eqn. 8 for 21 pairs of sources taking part in the case study for chemistry journals with the data sources A–G.

respect to d, there are certain pairs of sources with high, with medium and some with low compatibility. The compatibility between A–D and C–E is extremely high including 85 and 79% of the possible number of identical journals. Next are the sources C-D (62%) and A-C (54%), while the lowest compatibility occurs between the sources A–B and D–E (22%), and F–G (21%). From these results it follows that compatibility differs from one pair of sources to another independently of their size.

Based on these observations a hypothesis about possible compatibility impact was postulated: the effect of higher compatibility among the sources should be reflected on the quantity of the model output.

In an attempt to explore the impact of compatibility on the formation of the nucleus journals list, collections containing different number of data sources, were examined in terms of their average degree of compatibility ($\bar{d}$). Besides the case-study sample consisting of seven data sources ($\bar{d} = 42\%$, Fig. 4), the already exploited computer-simulated collections I–VI were considered. Two new collections were organized, each of them showing the absence of one constituent of the pair with the highest compatibility:

$n = 6$ VII A,B,C,E,F,G
VIII B,C,D,E,F,G.

For each of the nine collections, the number of journals in the nucleus and ($\bar{d}$) were correlated. The results are displayed in Table 6. Values for ($\bar{d}$), given in the first column in the increasing order, are found to vary from 35 to 47%. Further, a significant correlation between ($\bar{d}$) and N does exist (0.741), indicating that the degree of compatibility of the sources used is positively related to the actual number of journals constituting the nucleus.

Similar conclusion is reached if the effect of pairs of sources is analyzed. Let us start with the pair having the highest compatibility, A–D. The absence of this pair, as in the sets II, III, VII, and VIII, caused the number of nucleus journals to be rather low (302–329). On the other hand, the presence of these compatible sources resulted in the formation of the larger nuclei, N being in the range 340–389 as in IV–VI, for $n = 6$, and even reaching the top value of 463 in I, for $n = 5$. It is most probable that the compatibility of each pair of sources, present in a given set of data, plays a certain role leading eventually to the variations of N. However, the effect of the A-D pair always prevails. Namely, the presence/absence of either both sources A and D, or only one of them, would affect three, or at least two, of the highest positions on the compatibility scale shown in Fig. 4 (A–D, C–D, A–C, respectively). In other words, all this is clearly presented by the average degree of compatibility on the aggregate level.

Although the understanding of the factors governing the mechanism of the nucleus journals list formation is still deficient, our analysis offers evidence that compatibility of sources really demonstrates the level of interactions between the sources enabling, plausible predictions about the results of the journal selection procedure.

Table 6. Input/Output analysis: Comparison of average degree of compatibility ($\bar{d}$) and the number of journals in the nucleus (N) as determined by the model proposed

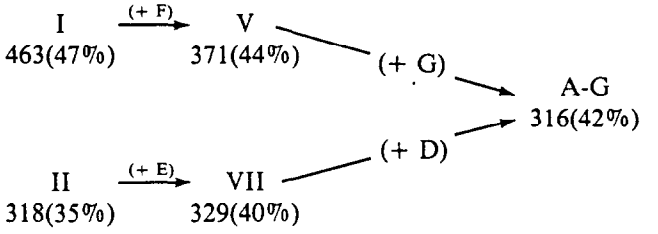
($\bar{d}$) (%)	Set	Sources excluded	N
35	II	D, E	318
38	III	C, D	320
40	VII	D	329
40	VIII	A	302
42	case-study	–	316
44	V	F	371
44	IV	G	389
45	VI	B	340
47	I	F, G	463

At this point, the impact of compatibility of sources should be considered in relation to:

1. the number of the chosen data sources, and
2. the concept of active combinations formulated by the model.

In the case of a small number of data sources ($n < 5$), high compatibility ($\bar{d}$ cca 45%) may act negatively due to an inappropriate increase of the number of journals which constitute the nucleus. On the contrary, low compatibility between the data sources ($\bar{d}$ less than 40%), especially in the cases where $n > 7$ are used, leads to a wider scatter of journals among the sources, resulting in a small number of journals in the nucleus.

The second relation could be explained by following the changes occurring in the course of the addition of new sources to the pool. There are two pathways to complete the case-study sample containing sources A–G, summarized in Scheme 3. The alterations in the number of nucleus journals and in average degree of compatibility (given in parentheses) are included (Scheme 3):



It was supposed [1] that, in shaping of the new nucleus by the addition of a data source, three groups could be recognized:

a. A group of journals present in the previous and in the new nucleus;
b. a group of journals in the new nucleus originating from the combinations that reached the level of activity by the action of the newly added source, and
c. a group which was present in the earlier nucleus, but due to the changes in the relative weights has now lost its activity and, consequently, it is eliminated from the new nucleus.

By applying this assumption to the examples in Scheme 3, it appears that the first group of journals, (a), might be identified as the content of higher ranked combinations, i.e., the corresponding reference overlappings in each of the sets (I, II, V, VII, and A–G).

Considerable diminishing of N that occurred in the first pathway is due to the predominance of (c) over (b). Stepwise addition of the sources F and G, both of rather low compatibility with all other sources, has not given rise to the formation of new active combinations, but on the other hand, it has contributed to the loss of activity of some combinations containing highly compatible pair A–D. The argument for such a statement could be found in Table 1: two- and three-fold overlappings (AD and ADE) are among the inactive combinations, the first of them has most probably lost its level of activity already in the process I → V.

6. CONCLUSION

The proposed journal selection model proved to be effective in the indirect evaluation of scientific journals. It provides an objective way to assess the nucleus of a periodicals collection for a particular discipline, subject to further supplementing. The conception of the model consists of:

1. involving several data sources of different types to make a pool,
2. ranking the sources chosen, and
3. evaluating the overlap on the basis of the active combinations.

The model's capability to cope with data sources of unequal quality should be pointed out. Even if data sources, considerably varying in quality and size, have to be used, these shortcomings would decrease through the processing, because the model itself acts restrictively and selectively.

Restrictive character of the model is incorporated into one of the criteria for ranking (indicator I_2), stressing the importance of the contents of the higher (reference) overlappings. By introducing such a restraining factor, quantities are made mutually consistent, with the results:

1. that the impact of larger sources is significantly reduced, and
2. that smaller sources are given a certain priority provided that they do contain journals in reference overlappings.

Although the quantities and proportions were somehow arbitrarily settled, they are internally consistent within the model system.

The selectivity of the model is manifested in the utilization of the concept of the active combinations. Verification study has shown that it operates as a suitable means for determination of the nucleus contents. Moreover, by fixing the range of the active combinations appropriately (a_c = 3.3–10.0), the model becomes generally applicable, regardless of the way of the pool construction. One of the advantages inherent to this journal selection model is its flexibility.

Acknowledgements—Thanks are due to Mr. I. Bárány and Mr. D. Kladarić for computer simulations, and to Dr. T. Radelja of the University of Split, for encouraging discussions. We are indebted to one of the referees for drawing our attention to the very pertinent references.

REFERENCES

1. Pravdić, N.; Oluić-Vuković, V. Application of overlapping technique in selection of scientific journals for a particular discipline—Methodological approach. Information Processing & Management, 23(1): 25–32; 1987.
2. Singleton. A. Journal ranking and selection: A review in physics. Journal of Documentation, 32(4): 258–289; 1976.
3. Rice, B.A. Selection and evaluation of chemistry periodicals. Science and Technology Libraries 4:43–59; 1983.
4. Evans, G.E. Collection evaluation: Developing library collections. Littleton, CO: Libraries Unlimited; 1979: 234–253.
5. Nisonger, T.E. An annotated bibliography of items relating to collection evaluation in academic libraries, 1969–1981. College and Research Libraries, 43(4): 300–311; 1982.
6. Kraft, D.H.; Hill, T.W. A journal selection model and its implication for a library system. Information Storage and Retrieval, 9(1): 1–11; 1973.
7. Robertson, B.E.; Hensman, S. Journal acquisition by libraries: Scatter and cost-effectiveness. Journal of Documentation, 31(4): 273–282; 1975.
8. Chudamani, K.S.; Shalini, R. Journal acquisition—cost-effectiveness of models. Information Processing & Management, 19(5): 307–311; 1983.
9. Kraft, D.H. Journal selection models: Past and present. Collection Management, 3(2-3): 163–185; 1979.
10. Rush, B.; Steinberg, S.; Kraft, D. Journal disposition decision policies. Journal of the American Society for Information Science, 25(4): 213–217; 1974.
11. Koenig, M.E.D. On-line collection analysis, Journal of the American Society for Information Science, 30(3): 148–153; 1979.
12. Broude, J. Journal deselection in an academic environment: a comparison of faculty and librarian choices. Serials Librarian, 3(2): 147–166; 1978.
13. Dhawan, M.; Phull, S.K.; Jain, S.P. Selection of scientific journals: A model. Journal of Documentation, 36(1): 24–32; 1980.
14. He, C.; Pao, M.L. A discipline-specific journal selection algorithm. Information Processing & Management 22(5): 405–416; 1986.
15. Science Citation Index, Institute for Scientific Information, Philadelphia, PA; 1982.
16. Chemical Abstracts Service Source Index, CAS, Columbus, OH; 1982.
17. ChemInform, VCH Verlagsgesellschaft mbH, Weinheim, FRG; 1982.
18. Eidgenossische Technische Hochschule, Zürich, Switzerland. Library Catalogue (microfishe): 1982.
19. List of Foreign Periodicals in Technical Libraries in Coordinated Library-Network. National Technical Information Center and Library, Vol. II, Budapest: pp. 63–90; 1982.

20. Pravdić, N.; Tóth, T.; Bárány, I. Selective dissemination of information as a basis for identification of the users' needs for scientific journals. Kemija u industriji, 34(6): 405–412; 1985 (in Croatian).
21. Pravdić, N.; Oluić-Vuković, V.; Tóth, T. Bibliometric analysis of contributions by scientists from Croatia (Yugoslavia) in the field of chemistry: Rank-frequency distribution. Kemija u industriji, 31(7): 351–356; 1982.
22. Pravdić, N.; Kritovac, D.; Aganović-Boras, A. Publications based on dissertations defended at the University of Zagreb in the period 1950–1980. Informatologia Yugoslavica, 19(3–4): 163–180; 1987.
23. Pravdić, N.; Aganović-Boras, A.; Kritovac, D. In search of a 'non-Citation Index' indicator for scientific activity assessment in less developed countries. Case study of Croatia/Yugoslavia. Scientometrics, 14(1–2): 111–125; 1988.
24. Oluić-Vuković, V. Impact of productivity increase on the distribution pattern of journals. Scientometrics, 17(1–2): 97–109; 1989.
25. Pravdić, N.; Oluić-Vuković, V. Dual approach to multiple authorship in the study of collaboration/scientific output relationship. Scientometrics, 10(5–6): 259–280; 1986.