Scientometrics, Vol. 3. No. 1 (1981) 55–63

THE ANALYSIS OF SQUARE MATRICES OF SCIENTOMETRIC TRANSACTIONS

D. de SOLLA PRICE

History of Science, Yale University, New Haven, Conn. 06520 (USA)

(Received April 22, 1980)

A method is explained for analysing square matrices of statistics giving transactions between each member of a set of nations, papers, journals, etc. In general self-transactions are different in kind to other exchanges of money, citations, etc., and a special method is given to compute row and column coefficients without relying on the diagonal elements. It is shown that this method yields very satisfactory analyses for journal and national citation data, enabling the members of the set to be assigned measures of size, quality and self-interest and a fuzzy set of clustered members from which all data may be derived.

An especially important type of matrix, common to the measurement of science, books, and money is the variety that registers transactions between all the members of some group. The matrix consists of a square array with the set of members arranged in the same order both in the rows and in the columns, and each cell then records the transaction from some member to some other.

Familiar examples are the import and export trade of the nations of the world, mail and telephone messages exchanged between cities and patents secured by nationals of one country in another. In our field perhaps the most important cases are with transactions in citations between nations, journals, authors, etc.

The essential difficulty in analysing all square matrices of this type is due to the fact that the "self-transactions" which should be given in the diagonal of the matrix are often undefined or constitute a physical process different than that governing the "foreign trade" of the members. For example the trade of a nation with itself or patents awarded to its own nationals may be recorded and, indeed, motivated from processes quite different from external relations. One therefore ought to ignore the evidence of the diagonal terms, in the first instance, when seeking to analyse such matrices in terms of row and column coefficients which determine the greater part of the data in all the other cells.

Fortunately it is possible to circumvent this difficulty by a straightforward procedure that replaces the diagonal terms by those that would be there if each member of the group treated itself in just the same way as it treats others. Once the diagonal has been reconstructed the matrix can be analysed in the normal way* as a product of row and column coefficients (giving the import and export strengths of the members) and an expectation factor matrix where the cells are all near unity for related members, high or low only in cases of unexpectedly strong or weak interaction in relatively few pairs of members, and zero for unrelated members.

The best procedure for reconstruction of a diagonal can be illustrated by considering the example of a 5 X 5 matrix in which each element is exactly a product of the corresponding row and column coefficients, but the diagonal terms function as unknowns.

Column coefficients

		f	g	h	i	j	Π_r
Row coefficients	a	•	ag	ah	ai	aj	$a^4 ghij$
	b	bf	•	bh	bi	bj	$b^4 fhij$
	c	cf	cg	•	ci	cj	$c^4 fgij$
	d	df	dg	dh	•	dj	$d^4 fghj$
	e	ef	eg	eh	ei	•	$e^4 fghi$
	Π_c	$f^4 bcde$	$g^4 acde$	$h^4 abde$	$i^4 abce$	$j^4 abcd$	$\Pi_r \Pi_c = (abcdefghij)^4$

If one forms the products of all the elements by row and by column, the grand product at the bottom right must contain each coefficient taken four times over. The fourth root of this grand product gives the product of all coefficients. Consider now the sub-matrix formed by deleting, for example, the first row and the first column. The grand product for the sub-matrix will then also be such that its third root (each coefficient now occurs only three times over!) is the product of all the involved coefficients bcdeghij. The original product divided by that now computed for the sub-matrix yields the product *af* which is the missing first term in the diagonal being reconstructed. All other diagonal terms may be calculated correspondingly. As an aid for the computation we may note that the grand product of the sub-matrix need not be calculated *ab initio* since it is given as the original grand product divided by the first row product and the first column product.

*See previous paper, "The Analysis of Scientometric Matrices for Policy Implications".

We have therefore

$$af = \sqrt[4]{\Pi_r \Pi_c} / \sqrt[3]{\Pi_r \Pi_c / \Pi_a \Pi_f} = \sqrt[3]{\Pi_a}\, \sqrt[3]{\Pi_f} / \sqrt[12]{\Pi_r \Pi_c}$$

and all the other diagonal elements may be reconstructed in the same fashion which uses all of the available data more than any other procedure.* If any cells of the matrix are vacant, unity may be inserted to replace the blank for this stage only.

With the diagonal elements reconstructed one may proceed to analyse the matrix as a product of row and column coefficients in the usual fashion. Summation by rows and by columns yields two sets of size factors and each element (including any real diagonal terms previously ignored) may have its expectation calculated as

$$\frac{\text{element} \times \text{grand total}}{\text{row sum} \times \text{column sum}}$$

As an illustration consider a given matrix:

$$\begin{array}{c|cccc|c}
 & & & & & \sqrt{\Pi_r} \\
 & (25) & 10 & 15 & 20 & 54.7 \\
 & 6 & (48) & 18 & 24 & 50.9 \\
 & 7 & 14 & (63) & 28 & 52.4 \\
 & 8 & 16 & 24 & (62) & 55.4 \\
\hline
\sqrt{\Pi_c} & 18.3 & 47.3 & 80.5 & 111.5 & 200.7 = \sqrt[3]{\Pi_r \Pi_c}
\end{array}$$

This becomes with reconstructed diagonal:

$$\begin{array}{c|cccc|c}
 & & & & & \Sigma_r \\
 & 5 & 10 & 15 & 20 & 50 \\
 & 6 & 12 & 18 & 24 & 60 \\
 & 7 & 14 & 21 & 28 & 70 \\
 & 8 & 16 & 24 & 32 & 80 \\
\hline
\Sigma_c & 26 & 52 & 78 & 104 & 260 = \Sigma_r \Sigma_c
\end{array}$$

The first step is to derive row and column coefficients by forming square roots of the row and column products, omitting the diagonal terms. Secondly, we find the

*For a similar procedure used in analysing international import/export matrices see (Deutsch & Savage).

cube root of the product of either the row or column coefficients. Thirdly, we reconstruct diagonal terms as the product of the appropriate row and column coefficients divided by the cube root of the grand product already derived; the terms are 4.98, 11.99, 21.02, 31.99 or within the accuracy of determination 5, 12, 21, 32. Fourthly, we form row and column sums of the reconstructed matrix. Last and fifthly we compute the row and column headings and expectation matrix that in this case is unity except in the diagonal and shows us that the entire matrix was formed from the data.

	Column Heading	Row Heading	Diagonal Strength
1	1	5	5
2	2	6	4
3	3	7	3
4	4	8	2

We now give a complete calculation of this type for the citation traffic between a set of eight journals important in biochemistry and listed in the 1977 *Journal Citation Index* (Table 1).

Following this procedure we reconstruct diagonal terms from the sixth roots of the row and column products and the seventh root of either set of such products. Using the reconstructed diagonals we now derive row and column sums and from these compute the ratio of actual to expected values. The results show that the given matrix may be computed from row, column, and diagonal strengths and from a residual matrix (Table 2) that is highly symmetric about the diagonal (the average difference between paired terms is ±0.12 and contains values that are mostly near unity (average deviation is ±0.29).

By averaging paired values of the expectation matrix we derive a new matrix (Table 3) showing the "fuzzy set" relations of the eight journals to each other. As with the general method for handling expectation matrices we need note only the data that are unexpectedly high or low. These are:

High Values	Pairs of Journals	Low Values	Pairs of Journals
2.09	3,5	0.50	6,7
1.77	3,7	0.52	3,6
1.71	5,7	0.64	1,5
1.57	2,6	0.70	2,3
		0.71	4,5

Table 1
Journal to journal data from 1977 *Journal Citation Index*

Journal	#	1	2	3	4	5	6	7	8	$\sqrt[6]{\Pi}$	Cited sum
J. Biol. Chem.	1	(9384)	6181	2107	3750	609	2335	719	2511	7075	22 764
Bio. Bio. Acta	2	2406	(7550)	865	1757	365	1378	408	1120	3072	10 806
Proc. N. A. S.	3	2770	2184	(3995)	1946	1470	488	1239	1329	4904	13 083
Biochem. U.S.	4	2553	2591	1057	(3827)	299	653	601	887	2999	10 183
Nature	5	1007	1230	1407	837	(2963)	379	603	630	2438	6 412
Biochem. J.	6	1183	1812	326	632	201	(2464)	150	528	1384	5 126
J. Mol. Bio.	7	1109	1136	1251	1347	504	216	(2545)	367	2096	6 234
Bio. Bio. R. C.	8	1624	1719	695	1040	263	564	241	(1313)	2040	6 766
$\sqrt[6]{\Pi}$		5760	7307	3026	4602	1173	1902	1297	2720	$\sqrt[7]{\Pi} = 8953$	
***New diagonal		4552	2507	1657	1542	319	294	304	620		
Citing sum		17 204	19 360	9362	12 851	4030	6307	4265	7992	$\Sigma\Sigma = 81\ 374$	
***quality		1.32	0.56	1.39	1.02	1.59	0.81	1.47	0.85		

Table 2
Expectation matrix for journal data

	1	2	3	4	5	6	7	8
1	(1.95)	1.14	0.80	1.04	0.54	1.32	0.60	1.12
2	1.05	(2.94)	0.70	1.03	0.68	1.65	0.72	1.06
3	1.00	0.70	(2.65)	0.94	2.27	0.48	1.81	1.03
4	1.19	1.07	0.90	(2.38)	0.59	0.83	1.13	0.89
5	0.74	0.81	1.91	0.83	(9.33)	0.76	1.79	1.00
6	1.09	1.49	0.55	0.78	0.79	(6.20)	0.56	1.05
7	0.84	0.77	1.74	1.37	1.63	0.45	(7.79)	0.60
8	1.14	1.07	0.89	0.97	0.78	1.08	0.68	(1.98)

Table 3
Symmetrized expectation matrix for journals
(average of a_{ij} and a_{ji})

	1	2	3	4	5	6	7	8
1		1.09	0.90	1.11	0.64	1.20	0.72	1.13
2			0.70	1.05	0.74	1.57	0.74	1.06
3				0.92	2.09	0.52	1.77	0.96
4					0.71	0.80	1.25	0.93
5						0.77	1.71	0.89
6							0.50	1.06
7								0.64

All other elements are closer to unity than the scattering due to random noise and we therefore derive the result that all eight journals are grouped together, but journals 3,5, and 7 form one tightly linked sub-cluster, and 2 and 6 form another cluster as little related as possible to the first. The remaining journals 1, 4, and 8 lie in the cluster, but distinct from the two sub-clusters generated.

What remains is a discussion of the row, column and diagonal strengths that give the greater part of the observed data. The output strength gives us a measure of general size for each current journal, and the input strength gives the size, as perceived from all other journals of its archive of published papers. The ratio of input to output sizes may be taken as a quality measure, but it must be remembered that one should not compare a new journal with a necessarily small archive with one that is old and large, furthermore a journal with many review articles will have anomalously large output size with its many references. Diagonal size as a multiple of expectation is more difficult to interpret. Most journals seem to show a multiple in the range 2.0–3.0, but in this case we have three journals with high values of 6.20, 7.79, and 9.33 respectively. What these values tell, I suppose, is that these particular journals cite themselves so much more than one would suppose from the rest of the group behavior because they each contain something like three times as much subject matter as the field defined by the remaining journals. Probably the best tactic with this parameter is to suppose that the outflow of references from each journal consists of a "domestic" or intra-journal portion (actual self-citation – expected self-citation) and a remaining portion which is inter-journal.

If one now took a much larger matrix containing the entire family of journals citing and cited by such broad journals as *Nature, Biochemical Journal,* and *Proceedings of the National Academy of Science* one would find an expectation matrix containing many more ones for those rows and columns and near zeros for the other more specialized journals. Ideally one should now take the grand matrix of more than 2000 journals in the *Journal Citation Index* and compute the output, input, and diagonal parameters and then have the symmetrized expectation matrix which would contain near-unity and near-zero values enabling all journals to be grouped and mapped together with their measures of size and "quality".

A similar analysis may also be made for the nation-by-nation citations for each field of science or for all literature in general. In illustration we take such a matrix (J. Davidson *Frame* and John J. *Baum,* "Cross-National Information Flows in Basic Research: Examples Taken from Physics", *JASIS,* 29, 5, September (1978), 247–252) for plasma physics (Table 4). The result of reconstructing the expected diagonal values and deriving row and column coefficients is an expectation matrix (Table 5) that is again strikingly symmetric. In this case moreover the input and output sizes (together with diagonal sizes reproduce successfully all the elements of the data ma-

Table 4
Narin – Plasma physics citations

Reconstructed matrix

	US	SS	JA	FR	WG	UK	EE	OT	Σ_r
US	(569.7)	190.3	80.3	73.0	76.5	46.4	26.3	186.6	1249.1
SS	284.4	(106.5)	48.5	34.6	45.5	17.7	19.7	93.8	650.7
JA	140.9	50.3	(21.8)	16.5	15.5	12.3	8.0	37.9	303.2
FR	112.8	36.6	17.5	(13.1)	11.9	9.6	6.5	31.7	239.7
WG	88.2	34.3	12.8	9.4	(9.9)	6.1	4.7	24.5	189.9
UK	78.5	22.1	12.4	9.2	7.6	(5.5)	4.4	21.9	161.6
EE	55.9	23.7	10.8	7.7	7.0	5.6	(3.6)	18.3	132.6
OT	264.7	85.7	35.6	27.9	28.4	19.2	12.9	(70.7)	545.1
Σ_c	1595.1	549.5	239.7	191.4	202.3	122.4	86.1	485.4	3471.9

Expectation matrix

US	(1.54)	0.96	0.93	1.06	1.05	1.05	0.85	1.07
SS	0.95	(4.09)	1.08	0.96	1.20	0.77	1.22	1.03
JA	1.01	1.05	(1.64)	0.99	0.88	1.15	1.06	0.89
FR	1.02	0.96	1.06	(1.00)	0.85	1.14	1.09	0.94
WG	1.01	1.14	0.98	0.90	(1.10)	0.91	1.00	0.92
UK	1.06	0.86	1.11	1.03	0.81	(1.32)	1.10	0.97
EE	0.92	1.13	1.18	1.05	0.91	1.20	(1.31)	0.99
OT	1.06	0.99	0.95	0.93	0.89	1.00	0.95	(0.93)

US – United States.
SS – Union of Soviet Socialist Republics.
JA – Japan.
FR – France.
WG – West-Germany.
UK – United Kingdom.
EE – Eastern Europe.
OT – Other.

trix to an average accuracy of only ±8% so that there is comparatively little "structure" in the relationships between the nations. Taking the highs and lows one finds:

High		Low	
1.17	SS,EE	0.82	UK,SS
1.17	SS,WG	0.86	UK,WG
1.15	EE,UK	0.87	WG,FR
1.13	UK,JA	0.88	EE,US
1.12	EE,JA		

Table 5
Symmetrized expectation matrix
(fuzzy set) for nations

US	SS	JA	FR	WG	UK	EE	OT	
	0.96	0.97	1.03	1.03	1.03	0.88	1.06	US
		1.06	0.96	1.17	0.82	1.17	1.01	SS
			1.02	0.93	1.13	1.12	0.92	JA
				0.87	1.08	1.07	0.93	FR
					0.86	0.95	0.90	WG
						1.15	0.98	UK
							0.97	EE

Mean deviation
1 ± 0.08
Mean asymmetry ±0.06

Country	(Citing) Output size	%	(Cited) Input size	%	Quality = Input/Output	Diagonal = Self-interest
US	1249.1	35.0	1595.1	45.9	1.28	1.54
SS	650.7	18.7	549.5	15.8	0.84	4.09
JA	303.2	8.7	239.7	6.9	0.79	1.64
FR	239.7	6.9	191.4	5.5	0.80	1.00
WG	189.9	5.5	202.3	5.8	1.07	1.10
UK	161.6	4.7	122.4	3.5	0.76	1.32
EE	132.6	3.8	86.1	2.5	0.65	1.31
OT	545.1	15.7	485.4	14.0	0.89	0.93

Of these only the expected grouping of USSR with Eastern Europe and the less expected similar strength grouping of USSR with Germany, and the dissimilarity between UK and USSR are striking. The single greatest irregularity is in the diagonal size 4.09 for the USSR. This implies that only the 106.5 of the actual 421.4 self-citations of the USSR are expected and more than 300 of the self-citations represent a domestic rather than international product in this field. This happens to be true in general for most fields in the USSR so that the input and output sizes for that nation are considerably less than would appear from number of papers published and references traded.

In short then, this simple model works very well, and shows that most square matrices of transactions that we meet in scientometric analysis can be reduced to a set of parameters, three for each entity. In the examples worked each nation or journal has an export size, an import size and some sort of self-interest (e.g. self-citation) coefficient and from these all entries in the matrix can be reconstructed, leaving behind a fuzzy set matrix showing the clustering, if any, of relationships between the entities engaged in the transactions. The analysis therefore yields measures of size, quality, self-interest and also a relational pattern of clustering. There is little doubt that this method provides a suitable analysis for the large *Journal Citation Index* data and also for bibliometric transactions between nations and even individual authors in particular fields or for aggregated data.