QUASI-CORRESPONDENCE ANALYSIS ON SCIENTOMETRIC TRANSACTION MATRICES

R. J. W. Tijssen, *J. de Leeuw & A. F. J. van Raan

*Science Studies Unit , LISBON-Institute / *Department of Data Theory*
University of Leiden, Stationsplein 242, 2312 AR Leiden
(The Netherlands)

In principle, a scientometric transaction matrix can be modelled by assuming that the number of transactions is the result of independent row and column contributions. More often one is primarily interested in the cross-structural relations between the participating entities, whereas the row and column marginals are of lesser or no importance. The values of the residuals after fitting an independence model to a complete transaction matrix can be analyzed by correspondence analysis to investigate the structure of the transactions between the rows and columns, after correcting for the marginals.

Recently a modification of correspondence analysis has been developed, quasi-correspondence analysis, which seems quite suitable for the analysis of citation-based transaction matrices which are incomplete or in which the incorporation of certain transactions may seem inappropriate. An illustration of both data analysis-techniques will be given using a journal-to-journal citation matrix.

Introduction

The rapidly increasing accumulation of scientific knowledge has been an important incentive in the development and application of so-called "science indicators", i.e. research methods to assess, categorize and measure specific characteristics of science, such as the effectiveness of scientific work and research performance. The so-called bibliometric indicators of science often use quantified characteristics of scientific literature as, for instance, the number of references within scientific publications. These citations to related publications can be seen as an recognition of such work.

The validity of a citation as such an elementary unit of communication between science entities (e.g., publishing entities like authors, journals or research groups/institutes) remains a matter of controversy : one must not only consider the existence of negative citations, irrelevant citations and self-citations, but also the possible citation-database limitions as well as existing (field and time-dependent) citation practices (cf. *Moed et al.*[1]). *Mac Roberts* and *Mac Roberts*[2] have also given evidence that relevant publications are often missing in the reference list and references, in general, do not accurately cover the topic(s) discussed. Consequently, extreme opinions are to be found on the use of citations in constructing quantitative measures of science, extending from those who would employ citation-based measures as a reliable indication of the peer recognition of scientific work to those who only advocate the use of citations in literature search. Citation-based measures have nevertheless become important science indicators because citation-data often yield relatively unobtrusive information on the relations between scientific entities, especially when one is dealing with citations of higher levels of aggregation. In practice such citation-based quantitative measures evaluating the "impact" of scientific activities in a given subfield of science, or between various subfields, have not only proved to be an valuable tool in assessing characteristics or even effectiveness of scientific activities (cf. *Garfield*[3]), but also for providing data in science policy studies (cf. *Moed et al.*[4-5]). In view of these pros and cons mentioned it will assumed in this paper that citations yield an indication of the scientific merit and of the utility of the cited object, and citation-transactions thus provide a useful measure to evaluate interrelationships between scientometric entities.

Aggregated transaction data are often displayed in a matrix, in which sets of entities are assigned to the rows and columns. Each cell of such a transaction-matrix contains citation-values indicating the level of transaction between a row and a column entity, in general the observed number of citations. When one is interested in the interrelationships between the same set of entities, the rows classify the cited-mode and the columns the citing-mode of an entity or vice versa. The elements in the main diagonal of such a matrix represent the self-transactions. In general, the number of self-transactions is (much) higher than the off-diagonal transaction-values. They are often the result of specific features in the transaction process. For example, in the case of journal-to-journal citation-transactions the self-citations within scientific journals is partly due to the fact that authors within a specific field tend to utilize a selection of the scientific journals covering that field. This mechanism can result in a high concentration of citations to publications in the journal in which both the cited author as well as the citing author publish.

There are also a number of data analysis-techniques, in which the structure of the transaction matrix can be investigated by modelling the citation-data. An adequate model of the proportionally high number of self-citations will generally lead to a non-adequate modelling of the other elements in the matrix and thus obsure the off-diagonal relationships. Since our interest is mainly on the (off-diagonal) citation-relations between entities these diagonal elements are of less or no importance in the modelling-process. In order to discard these self-transactions one can proceed by adjusting or eliminating these values to minimize their effect on the analysis results.

In this paper such a data analysis-technique is discussed yielding information on the relations between entities, of both the row and column-mode, while ignoring (the values in) specified matrix elements. Such elements may be ignored because they are unobserved or missing, because they pertain events that cannot occur, or because they are unusual in other respects. We discuss the case in which the self-citations of a journal-to-journal citation matrix are adjusted, but the formalism to be developed is completely general and can be applied to any element or a group of elements in any rectangular (scientometric) matrix.

Modelling the transaction data

A large influence of one entity on another entity or a strong interrelationship between scientometric entities can be expressed by a relatively high number of citation-transactions. In general, one can consider the strength of a relationship as a function of the number of these transactions. Scientometric transaction matrices with entities from different domains will generally contain a transaction structure quite different from the expected structure if one only considers the marginals as in the case of the so-called independence model. In this model one assumes that the existing transaction levels between the entities are only determined by the "size" of the entities : the rows i (i=1,..,i',..,I) and columns j (j=1,..,j',..,J) have an independent contribution to the cell-frequencies i.e. the probability of a transaction from object i to object j is equal to the probability of a transaction from column i multiplied by the probability to receive a transaction in row j, thus in terms of the population parameters π

(1) $\pi_{ij} = \pi_{i+} \pi_{+j}$

with the '+' de..oting the summation over the omitted index.

It is well known that the maximum likelihood estimator (MLE) of the expected number of transactions, obtained from the datamatrix with the observed number of transactions $X=\{x_{ij}\}$ is equal to

(2) $e_{ij} = x_{i+} \, x_{+j} / n$

with $x_{++} = n$

The significance of the difference between the observed value and expected values derived from this model can be evaluated with the Pearson chi-square statistic:

(3) $$\chi^2 = \sum_i \sum_j (x_{ij} - e_{ij})^2 / e_{ij}$$

This statistic has an asymptotic chi-square distribution with (I-1) (J-1) degrees of freedom (df). If the value of χ^2 has a probability near zero according to the chi-square distribution, the expected and observed transaction values are significantly different. In such cases one might consider an alternative model with additional parameters to account for the variation in the citation frequencies due to first- and/or higher-order interactions between rows and columns of the matrix.

Of course the independence model is often only a baseline-model and it is obvious that the expected values based on this model will generally not fit the citation-data. The differences between the original citation data and the expected values will yield useful information on the citation-relations, because the "size-effects" will then have been ruled out. Since we are primarily interested in an analysis of residuals after fitting a suitable restrictive model we focus on the independence model. More sophisticated models (i.e. with more model-parameters) will generally yield a better fit of the data, but often at the cost of problems when interpreting the multitude of parameters and often leaving non-informative residuals. A serious drawback prevents fitting the independence model to a transaction matrix without, for example, involving self-transactions: the ML-estimates of the values in the off-diagonal elements of the matrix cannot be computed directly without the main diagonal. A solution to this problem can be found by introducing the quasi-independence model ; a generalisation of the independence model to incomplete matrices.

Quasi-independence

The quasi-independence model enables one to ignore elements in the matrix and still fit an independence model on the remaining observation values {(i,j) in a given subset of the index pairs L} based on a MLE-procedure, while estimating the expected values for observations which are not modeled {(i,j) not in L}. A brief outline of the quasi-independence model will be given for a two-way matrix. For a more detailed discussion of the concepts the reader is referred to *Goodman*[6].

Contrary to the independence model, a direct estimation of the expected frequencies is impossible when one fits a model based on quasi-independence. In this case an iterative maximum likelihood algorithm of the following 'iterative proportional fitting' type (cf. *Deming&Stephan* [7]) can be applied : if the matrix with the observed frequencies x_{ij} is complete, one can fit a quasi-independence model π which assumes that $\pi_{ij} = \alpha_i \beta_j$ for all (i,j) in L. The π_{ij} with (i,j) not in L are unrestricted and not estimated ; these values are found by substituting the observed values, thus $\pi_{ij} = x_{ij}$. The multinomial likelihood equations are

(4a) $\sum_j \{x_{ij} \mid j \in J(i) = \sum_j \{ \pi_{ij} \mid j \in J(i)\}$

(4b) $\sum_i \{x_{ij} \mid i \in I(j) = \sum_i \{ \pi_{ij} \mid i \in I(j)\}$

with I(j) and J(i) respectively denoting the index i and the index j for which the cells (i,j) are in L. It follows from eqs. 4a,4b that the expected marginals of the restricted cells have to be equal to the observed marginals.

The quasi-independence algorithm starts by setting the elements (i,j) in L equal to the parameter product $\alpha_i\beta_j$, for an arbitrary choice of α and β. A convenient choice is $\alpha_i \beta_j=1$. For each i all elements π_{ij} with $j \in J(i)$ are multiplied with an constant, satisfying eq. 4a. Only the row-sums add up to the correct marginal numbers. Subsequently the same procedure is applied to the columns j with $i \in I(j)$, resulting in correct column-sums, but undoing the correct row-sums. This process is repeated until all values π_{ij} converge to stable values with an acceptable level of accuracy.

Price [8] made a first attempt to fit an quasi-independence model on a square transaction matrix. In *Price's* procedure the diagonal elements are considered missing and initial values are assigned to the diagonal elements via a multiplicative model based only on the off-diagonal elements. The transaction matrix with the estimated self-transactions is subsequently used to compute the final estimates of both the diagonal and off-diagonal elements based on the independence model. A major shortcoming of this two-step method is the fact that it lacks a proper conceptual basis and each step leads to different expected values of the diagonal elements. *Noma* [9] elaborated on this procedure by fitting a quasi-independence model and subsequently a quasi-symmetric model. The latter model is an extension of the quasi-independence model in which an additional interaction-parameter is incorporated for each row and column-combination. The values of the interaction-parameters were used as input in a multidimensional scaling technique to compute a spatial indicator of similarity between the objects. *Noma's* use of the quasi-independence model and the quasi-symmetric model proved to be a conceptual improvement when modelling transaction data with dominating diagonal values. However, two critical aspects of this procedure to represent the relations between objects based on quasi-independence are to be considered. First, the spatial results clearly depend on the extent to which the data are fitted by the model. Using the parameter values of model with a lack of fit can only result in a rough approximation of the similarities between the entities involved. A more promising approach would be to fit a (highly) restrictive model, such as the quasi-independence model, and analyze the residuals to investigate the remaining structure between the entities. The residuals contain the information on the relations between the entities after correcting for the "size" of the entities and the size of the unproportionally high numbers of self-citations. Secondly, and more important, the rows and columns are treated symmetrically, whereas a transaction matrix is often highly asymmetric, reflecting large differences between row and column-mode of an entity. For example, the cited and citing characteristics between two journals can be of an entirely different nature, because the journals emphasize on different features of scientific reseach within a (sub)field, e.g. applied research versus basic research. This asymmetry is not accounted for in the symmetric model and remains hidden in the pattern of the residuals. These criticisms can however be dealt with by the data analysis-techniques : correspondence analysis and, more in particular, quasi-correspondence analysis.

Correspondence Analysis

Quasi-correspondence analysis (abbreviated to QCA) can be used to fit a quasi-independence model to a square transaction matrix and subsequently investigate the relations between the residuals in terms of the relations between scores assigned to the rows and columns. This technique is a generalisation of correspondence analysis (CA in the following), which is basically a standard eigenvector-eigenvalue decomposition of the matrix of residuals after fitting an the independence model. In short, CA can thus be seen as a technique which analyses a structure of values after correcting for the marginal frequencies. It can therefore be used complementary to loglinear modelling (cf.*Van der Heijden & De Leeuw* [10]), but it is also possible to interpret CA as a technique able to find a multidimensional representation of the dependence between rows and columns (cf. *Benzecri* [11]). The CA-results can be displayed in a simultaneous spatial representation of scores assigned to the rows and columns of a matrix.

Before describing QCA, a brief discussion of CA must be given. CA can be defined in terms of deviations from the independence model. Let X be the matrix with the observed number of standardized citation-transactions, with entries x_{ij} adding up to one. The row marginals x_{i+} are contained in the diagonal matrix D_r, and D_c contains the column marginals x_{+j}. The vector t contains elements equal to one. The matrix E is equal to $D_r\, t\, t'\, D_c/n$ with elements the e_{ij}. The '-sign denotes the transpose of a vector or matrix. The matrix containing the standardized residuals after accounting for the row and column effects is decomposed by computing the singular value decomposition

$$(5) \quad D_r^{-1/2}\,(X-E)\,D_c^{-1/2} = U\Omega V'$$

where U'U=I and V'V=I. Ω is a diagonal matrix containing the descending singular values ω_s, where s (s=1,..,p) is the index for the orthogonal solutions or, in geometrical terms, the so-called dimensions. If the independence model fits well, the residuals (X-E) are small resulting in small singular values. In this case, the CA-results are obviously of not much value ; the row and column parameters of the independence model are sufficient to approximate the number of citation transactions.

If one uses a limited number of CA-dimensions ($s<<p$) to describe the residual-structure each row and column is quantified by s quantifications using the corresponding elements of the eigenvectors. The resulting row and column scores are normalised by

(6a) $R = D_r^{-1/2}\ U\ n^{1/2}$

(6b) $C = D_c^{-1/2}\ V\ n^{\ 1/2}$

with a weighted average equal to 0 and a weighted variance equal to 1. Futhermore, $R'\ D_rR=nI$, $C'\ D_cC=nI$, $t'D_rR=0$ and $t'D_cC=0$.

The row and column scores are normalised in such a way that the Euclidean distance between a row i and a row i' of $R^*=R\Omega$ is equal to the chi-squared distance δ^2, which is defined as the distance between the respective row/column-profiles, were e.g. the profile of a column j is the column of the values x_{ij}/x_{+j}:

(7) $\delta^2(i,i') = (I_i - I_{i'})'\ D_r^{-1}\ X'\ D_c^{-1}\ X'\ D_r^{-1}\ (I_i - I_{i'})'\ n = (r_i\ - r_{i'})'\ \Omega^2\ (r_i - r_{i'}) =$
$(r^*_i - r^*_{i'})'(r^*_i - r^*_{i'})$

I_i and $I_{i'}$ are unit vectors from the identity matrix I. Approximations of the chi-square distances are found by only considering the columns of R^* corresponding to the p largest singular values of Ω. A similar approximation of the columns can be given by the Euclidean distances between the columns of $C^*=C\Omega$.

To facilitate the interpretation one can integrate the separate plots of the row and column scores into a single plot. This is done with the aid of a centroid principle expressed in a so-called transition formula, either

(8a) $R\Omega = D_r^{-1}\ X\ C$

or

(8b) $C\ \Omega= D_c^{-1}\ X'\ R$

Depending on the choice of the centroid principle the distances between a row point i and a column point j in such a joint plot can be interpreted by regarding the row points as the weighted average - or centroids - of the column points or vice versa. The row and column profiles D_r and D_c are used as weights.

If one interprets the relations between row scores (or column scores) using the chi-square distances, one must bear in mind the fact that rows or columns with similar profiles will have small distances between them, whereas large distances indicate considerably different profiles. The profiles of the marginal frequencies of X are always located in the origin of the plot ; points near the origin correspond to profiles resembling the mean profile. Row and column scores with profiles very deviant from the mean-profile, adding significantly to the chi-square total (cf. eq. 3), are found in the periphery of the plot. Using the transition formulas one can roughly interpret the distance between a row i and a column j ; the distance between the points is small if $x_{ij}>>e_{ij}$, points are far apart if $x_{ij}<<e_{ij}$.

The relation between the row and column scores and the original data is found through substituting eq. 6(a,b) in eq. 5, obtaining the so-called reconstitution formula:

$$D_r^{-1} (X-E)\, D_c^{-1} n = R\Omega C'$$

leading to

$$(9) \quad X = E + D_r R\, \Omega\, C'\, D_c\, n^{-1}$$

which clearly shows that correspondence analysis decomposes the departure from independence. The elements of $R\Omega C'$ are equal to $(x_{ij} - e_{ij})/e_{ij}$.

The Pearson-statistic χ^2 can be defined in terms of the so-called total 'inertia' of a CA-solution :

$$(10) \quad \mathrm{tr}\, \Omega^2 = \sum_s \omega^2_s = \chi^2/n \qquad \text{(tr=trace ; the sum of the diagonal elements)}$$

The importance of a dimension can now be interpreted as the ratio of the inertia in a dimension and the total inertia $\omega^2_s/\mathrm{tr}\, \Omega^2$, or more simply, as the proportion of χ^2, which is decomposed in a dimension s.

It has already been pointed out that the row and column scores can be represented in different ways as coordinates in a joint plot, each method with its own specific advantages and disadvantages (cf. *Van der Heijden* [12]). Using the centroid principle one can choose between a plot of $(R, C\Omega)$ or $(R\Omega, C)$. In both cases distances between row and column scores approximate chi-squared distances (cf. eq. 7). A symmetric interpretation can be applied by constructing a joint plot of $(R\Omega^{1/2}, C\Omega^{1/2})$, which is an approximation of the two centroid-representations. The approximation will be better as the values of the elements of Ω become less different. This option will thus spread the distortion of the approximated chi-squared distances equally over the rows and columns, but a clear interpretation of the results in terms of chi-square distances or the centroid principle is lost. The residuals however, can be represented in terms of scalar products between the row and column vectors; the nature and strenght of the spatial relationship between coordinates is determined by the length of the vectors from the origin to the points and the angle between them : e.g. a small angle between relatively long vectors indicates a strong relationship between the corresponding objects, whereas orthogonal vectors indicate unrelated rows or columns.

Quasi-Correspondence analysis

Generalizing CA to a technique capable of decomposing residuals after fitting a quasi-independence model is now quite straightforward. It will only be discussed briefly. Detailed information on the maximum likelihood algorithm and other specifics of the technique can be found in *De Leeuw & Van der Heijden* [13]. Starting from the datamatrix X a matrix Y is computed containing the maximum likelihood estimates under the quasi-independence model. The marginals D_r and D_c of X and Y are identical. The values of the non-modelled elements after fitting the quasi-independence model are similar to those in X, hence these elements have residuals equal to zero. Analogous to eq. 8 the singular value decomposition on the matrix of the residuals (X-Y) is computed :

(11) $D_r^{-1/2}(X - Y)D_c^{-1/2} = U\Omega V'$

The centroid principles are now written as

(12a) $R\Omega = D_r^{-1}XC - D_r^{-1}YC$

(12b) $C\Omega = D_c^{-1}X'R - D_c^{-1}YR$

The relation between the chi-square distances and the singular values (cf. eq. 3, eq. 10) is now lost because the trace of the singular values is equal to

$$(13)\quad \sum_i \sum_j (x_{ij} - y_{ij})^2 / d^r_i\, d^c_j$$

with d^r_i and d^c_j denoting the corresponding elements i and j in the diagonal matrices D_r and D_c.

A remedy for this unpleasent feature can be found by replacing the elements in the diagonal weighting matrices D_r and D_c (eq.10) with maximum likelihood estimators (α_i,β_j) of the corresponding rows i and columns j. In this case the trace of the singular values results in the familiar chi-square statistic for testing quasi-independence:

$$(14)\quad \sum_i \sum_j \{(x_{ij} - y_{ij})^2 / y_{ij} \mid (i,j) \text{ in } L\}$$

The centroid principles of this normalisation can be found as follows : suppose P is a matrix with elements r_{ij} equal to x_{ij} for all (i,j) in L, while r_{ij} is equal to $\alpha_i\beta_j$ for all (i,j) not in L. Note that this treatment of the elements is the reverse of the previous one : the quasi-independence algorithm now iterates on the diagonal elements. It converges to the same point as the iterative proportional fitting-algorithm in which both algorithms yield the same value for the chi-square test of quasi-independence. Let Q = Puu'P / u'Pu and u is a vector with unit-elements. If P is the matrix of observed values, Q is the matrix of expected values based on the independence model. In this case P-Q = X-E. The centroid principles are now be defined as

$$(15a)\quad R\Omega = D_r^{-1}PC$$

$$(15b)\quad C\Omega = D_c^{-1}P'R$$

Comparing these centroid principles with eq. 8 shows that quasi-correspondence analysis of the matrices X and Y, with weights d^r_i and d^c_j is identical to CA of the matrix P.

The characteristics of the different options to plot the row and column scores mentioned in the foregoing, also apply to the QCA-results. For reasons of interpretability the symmetric joint plot ($R\Omega^{1/2}, C\Omega^{1/2}$) of the CA and QCA-results is used in the following application to a journal-to-journal transaction matrix.

Application to astrophysical and astronomical journals

The data consist of citation counts between seven highly cited scientific journals from the United States and Europe, with publications on topics in the fields of astronomy and astrophysics. These two specific scientific subfields were chosen because journal articles are a predominant form of scientific communication within these (sub)fields. All journals used mainly consist of 'normal' articles ; letters or other types of short publications often appear in supplements or separate journals. The citation counts were collected by manual search from the 1983 *Journal Citation Reports* [15]. The result is given in table 1.

Table 1

Journal-to-journal data from the 1983 *Journal Citation Reports*

		Citing journal						
Cited journal		1	2	3	4	5	6	7
1	AA	2714	2009	867	296	297	129	159
2	AP	1506	16079	3383	1358	1186	895	975
3	MN	1163	2327	1959	268	315	175	141
4	AN	454	965	433	651	100	126	58
5	APSS	424	978	261	191	464	94	33
6	PASP	282	576	208	170	69	183	63
7	ARAA	237	601	140	64	74	32	45

An a priori differentiation between the journals is already apparent from the journal titles : two US journals, Astronomical Journal (AN) and Proceedings of the Astronomical Society of the Pacific (PASP), and the British journal Monthly Notices of the Royal Astronomical Society (MN) emphasize astronomical subjects, while the

US journal Astrophysical Journal (AP) concentrates on astrophysical topics. The European journal Astronomy and Astrophysics (AA) covers both fields. In addition to the publications on astrophysics, the US journal Astrophysics and Space Science (APSS) also contains publications on space physics and related topics on the solar system. The Annual Review of Astronomy and Astrophysics (ARAA) contains papers in which an overview is given of the past and current developments in various subfields of astronomy and astrophysics; these papers generally contain of large amount of references. Due to summarizing characteristics, the review papers are often highly cited by papers which deal with topics within the reviewed subfield(s).

The elements in the main diagonal of the non-review journals are dominating ; 44.9% of the total amount of citations are self-citations. In terms of the citing (column) totals the AP has the largest proportion of self-citations, namely 68.3%. The PASP has the lowest selfciting percentage with 11.0%. Obviously, the review characteristics of ARAA prevent a high self-citation rate ; the self-citations only amount to 3.1%. The values of the self-citations of the non-review journals all exceed the expected values on the basis of independent row and column contributions (cf. table 1 and table 2).

Table 2

Expected number of self-citations based on the independence model (rounded numbers)

Journal	AA	AP	MN	AN	APSS	PASP	ARAA
Expected value	1287	13583	936	170	125	52	36

Before we analyze the matrix with quasi-correspondence analysis it is illustrative to show the effect of those high diagonal values by computing a CA-solution on the complete matrix. A 4-dimensional solution was computed with the CA-program ANACOR (*Gifi* [15]). The chi-square equal to 7393.6 (df=36 ; p<.001), with singular values $\omega_1-\omega_4$ are equal to .056, .039, .028 and .021, respectively. The magnitude of the first and second singular value compared to the third and fourth singular value suggest a display of the row and column scores of first two dimensions as a parsimonious representation of the results with a relatively small loss of information. The two-dimensional solution decomposes 53% of the inertia ; 37% in the first dimension and 26% in the second dimension. The plot of the row- and column scores of the CA-solution is given as figure 1. As a result of the dominance of the diagonal elements over the off-diagonal elements the row and column profiles tend to become more similar and, consequently, relatively little differentiation is found between the row and column scores. In the case of the AP, AN and MN the row and column points are located very close to each other as a result of the proportionally high number of self-citations within these journal.

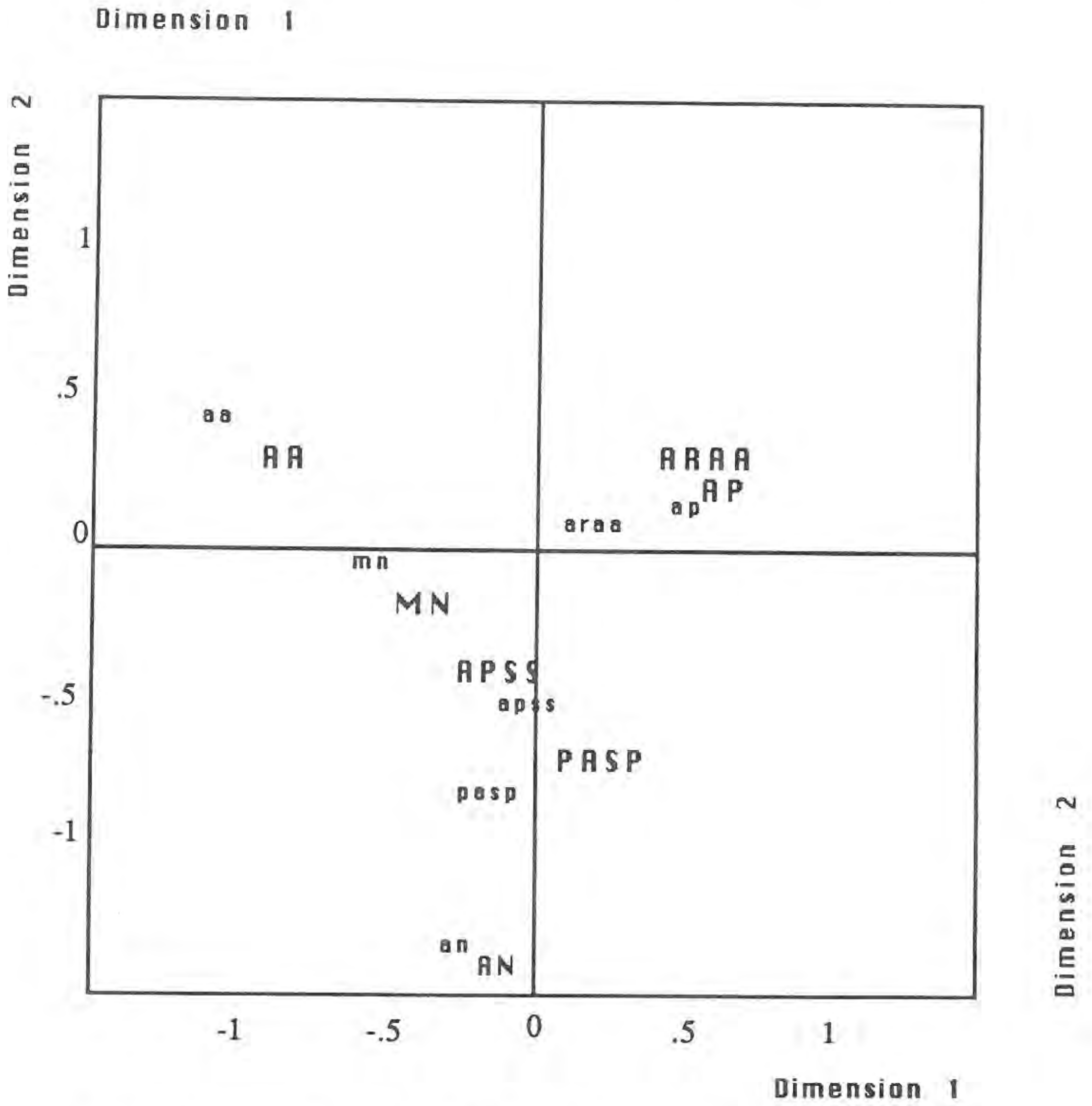


Fig. 1. Correspondence analysis of the journal-to-journal citations from the 1983 *Journal Citation Reports*. Large label=citing journal ; small label=cited journal. Labels are centered at the location of the journal-mode.

The first dimension of the CA-results - the horizontal axis - reveals a relationship which accounts for the largest amount of variance found between the journals: the duality between AP and ARAA (which have a strong citation-relationship - cf. table 2) and AA. The journals AA and AP thus have a weaker citation-interrelationship than would be expected, after correcting for the large row- and column sums for both journals. ARAA cites AP much more then expected, whereas the cited ARAA is located near the centre indicating a mean cited pattern i.e. the other journals cite ARAA in a column-proportional manner. Considering the contents of a review-journal such a result is likely to occur.

If one projects the scores of the other journals on the axes of the first dimension these journals have an intermediate position. The second dimension - the vertical axis - is used mainly to separate AN from the other journals.

Considering the contents of the journals, an overall interpretation seems to lead to the conclusion that the position of the three journals in the centre of the triangle (MN, APSS and PASP) is not so much the result of similarities between these journals, but a result of the differentiation between the US and European journals (in particular AA versus AP) in the first dimension, whereas the second dimension tends to differentiate between the (astro)physics-oriented journals (in particular AN) and the other journals.

The deviant values of the self-citations of the non-review journals clearly suggest a quasi-correspondence analysis on the data-matrix. The following QCA-analysis was computed with the use of a program written in APL. The display of the resulting row and column scores is given in figure 2.

Of course the remaining chi-square after fitting the quasi-independence model to the data is lower then in the case of the independence model (χ^2=354.7 ; df=36 ; p<.0001). However, this model still doesn't fit the data adequately The remaining inertia is decomposed with a two-dimensional QCA-solution with singular values equal to .062 and .049, accounting for respectively 42% and 26% of the inertia. A two-dimensional solution was mainly chosen to simplify the interpretation of the results. The three additional singular values of a 5-dimensional solution (.041, .033 and .012, respectively) would have justified a 3-dimensional or even a 4-dimensional solution.

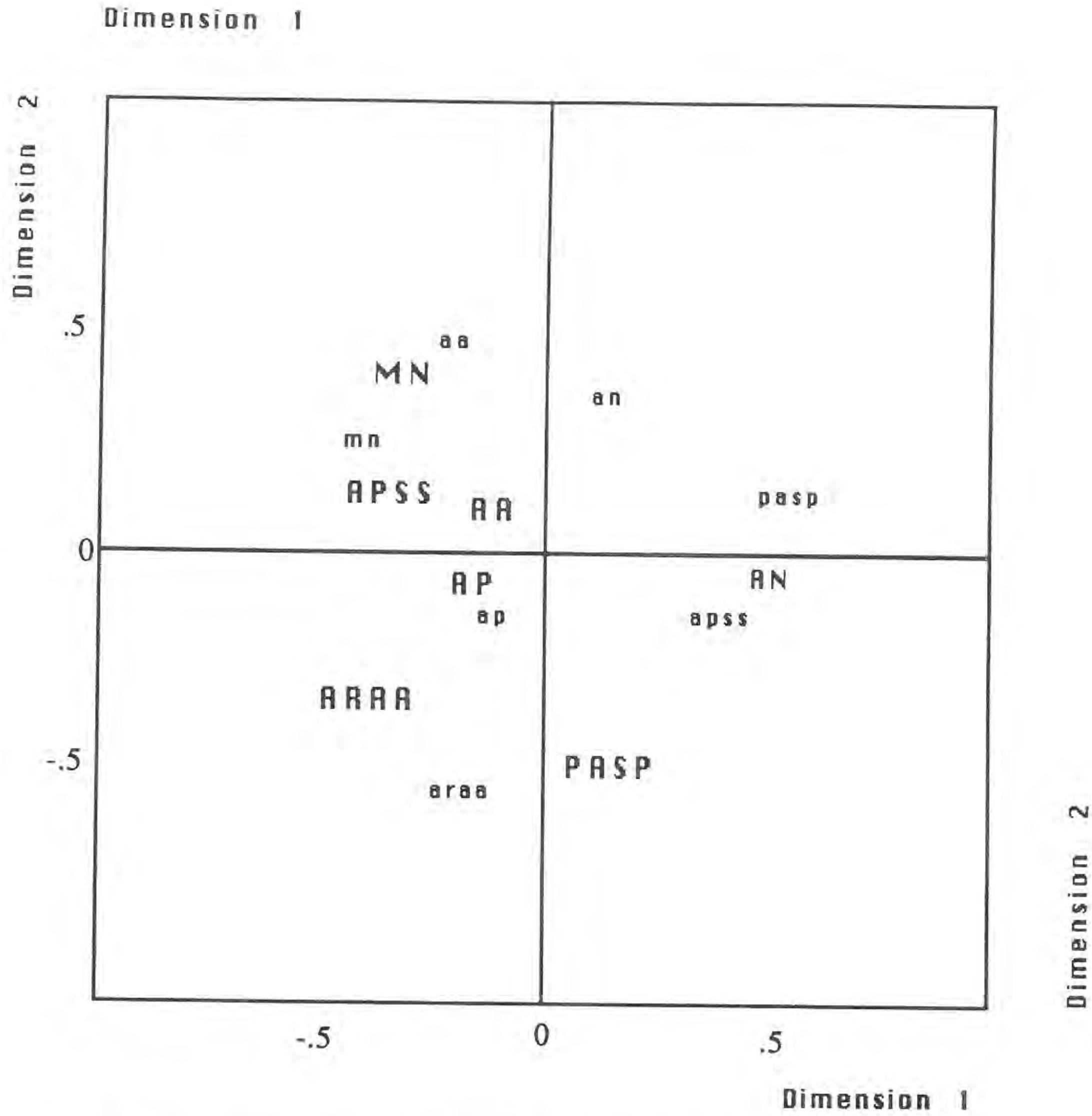


Fig. 2. Quasi-correspondence analysis of the journal-to-journal citations from the 1983 *Journal Citation Reports.* Large label=citing journal ; small label=cited journal. Labels are centered at the location of the journal-mode.

Elimating the effect of the large diagonal values has resulted in a drastic change in the display of the structure of transactions between the journals: the positions of the citing and cited modes of the journals are now no longer located near each other. These positions of the rows and columns yield a more accurate reflection of the structure of the journal-interrelations. Although the clear-cut differentiation between the separate journals from the CA-solution is lost, the row and column scores of the QCA-solution still display an overall structure which is still comparable with the CA-results, but after a 90-degree rotation of the axes. The distinction between the astronomical-oriented journals (AN, PASP and APSS) and the other more physics-oriented journals has now become the most important feature within the structure of residuals (table 3), after fitting a quasi-independence model. The rotation is specifically caused by the elimination of the self-citation values of the two highest (self-)cited journals : AP and AA. Without the self-citations the profiles of these journals are less dominating in the transaction-structure. The more prominent citing relationships of AN now become the most important feature in the first dimension.

This leads to a number of differences between the solutions: For example, the first dimension now focusses on the relationship between AN and PASP, APSS. Notice that PASP is cited more often by AN then expected, whereas in the reverse citation-process this is not the case. In fact, the orthogonality of the cited and citing modes of AN and PASP in the structure of figure 2 indicate that these citation-processes are relatively unrelated to each other. All scores except ARAA are located in a vertical band through the centre of the plot : the high citing journals AP, AA and MN have lost their peripheral position - their adjusted profiles now have a larger resemblance with the marginal profiles. AP takes a more central position in the plot, indicating that the other journals refer to AP in more or less proportional way and in their turn are also proportionally cited by AP. Both AA an APSS have a central position as citers, spreading their references more or less proportionally over all journals, with AA having a slight emphasis on MN.

Table 3

The standarized residuals after fitting the quasi-independence model

		Citing journal						
Cited journal		1	2	3	4	5	6	7
1	AA	.0	-.008	+.023	-.002	+.011	-.006	-.022
2	AP	-.003	.0	-.006	-.007	-.005	+.028	+.005
3	MN	+.013	+.004	.0	-.018	+.008	-.019	-.011
4	AN	-.004	-.006	+.020	.0	-.012	-.015	+.019
5	APSS	-.002	+.009	-.019	+.025	.0	-.026	+.006
6	PASP	-.002	-.008	-.004	+.043	-.005	.0	-.014
7	REV	-.005	+.014	-.019	-.008	+.004	+.011	.0

On the lower side of this band, the second dimension displays the relationship between theciting PASP and the cited ARAA; inspection of the matrix of standarized residuals reveals a relatively high positive citation-excess between these modes of PASP to ARAA. The strong CA-relationship between AP and ARAA is still visible in the QCA-solution, but is now considerably weaker. In the upper part of the band the relatively strong citing relationship between the European journals MN and AA still exists, especially for the citations from MN to AA. Clearly both modes of ARAA have also lost their core-position, which is surprising particularly considering the cited characteristics of such a journal. It turns out that ARAA takes a position of its own, largely determining the third QCA-dimension.

References

1 H.F. Moed, W.J.M. Burger, J.G. Frankfort & A.F.J. van Raan, The application of bibliometric indicators : Important field-and time-dependent factors to be considered, *Scientometrics* 8 (1985) 177-203.

2 M.H. Mac Roberts & B.R. Mac Roberts, Quantitative Measures of Communication in Science : A Study of the Formal Level. *Social Studies of Science* 16 (1986) 151-172.

3 E. Garfield, *Citation Indexing*, New York. Wiley-Interscience, 1979.

4 H.F. Moed, W.J.M. Burger, J.G. Frankfort & A.F.J. van Raan, On the measurement of research performance : the use of bibliometric indicators. Research Policy and Science Studies Unit, University of Leiden, 1983.

5 H.F. Moed, W.J.M. Burger, J.G. Frankfort & A.F.J. van Raan, The use of bibliometric data for the measurement of university research performance, *Research Policy* 14 (1985) 131-149.

6 L.A. Goodman, The analysis of cross-classified data : independence, quasi-independence, and interactions in contingency tables with or without missing entries, *Journal of American Statistical Association* 63 (1968) 1091-1131.

7 W.E. Deming & F.F. Stephan, On the least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Statist.* 11 (1940) 427-444.

8 D.J. Price, The analysis of square matrices of scientometric transactions. *Scientometrics* 3 (1981) 55-63.

9 E. Noma, An improved method for analyzing square scientometric transaction matrices. *Scientometrics* 4 (1982) 297-316.

10 P.G.M.van der Heijden & J. de Leeuw, Correspondence analysis used complementary to loglinear analysis. *Psychometrika* 50 (1985) 429-447.

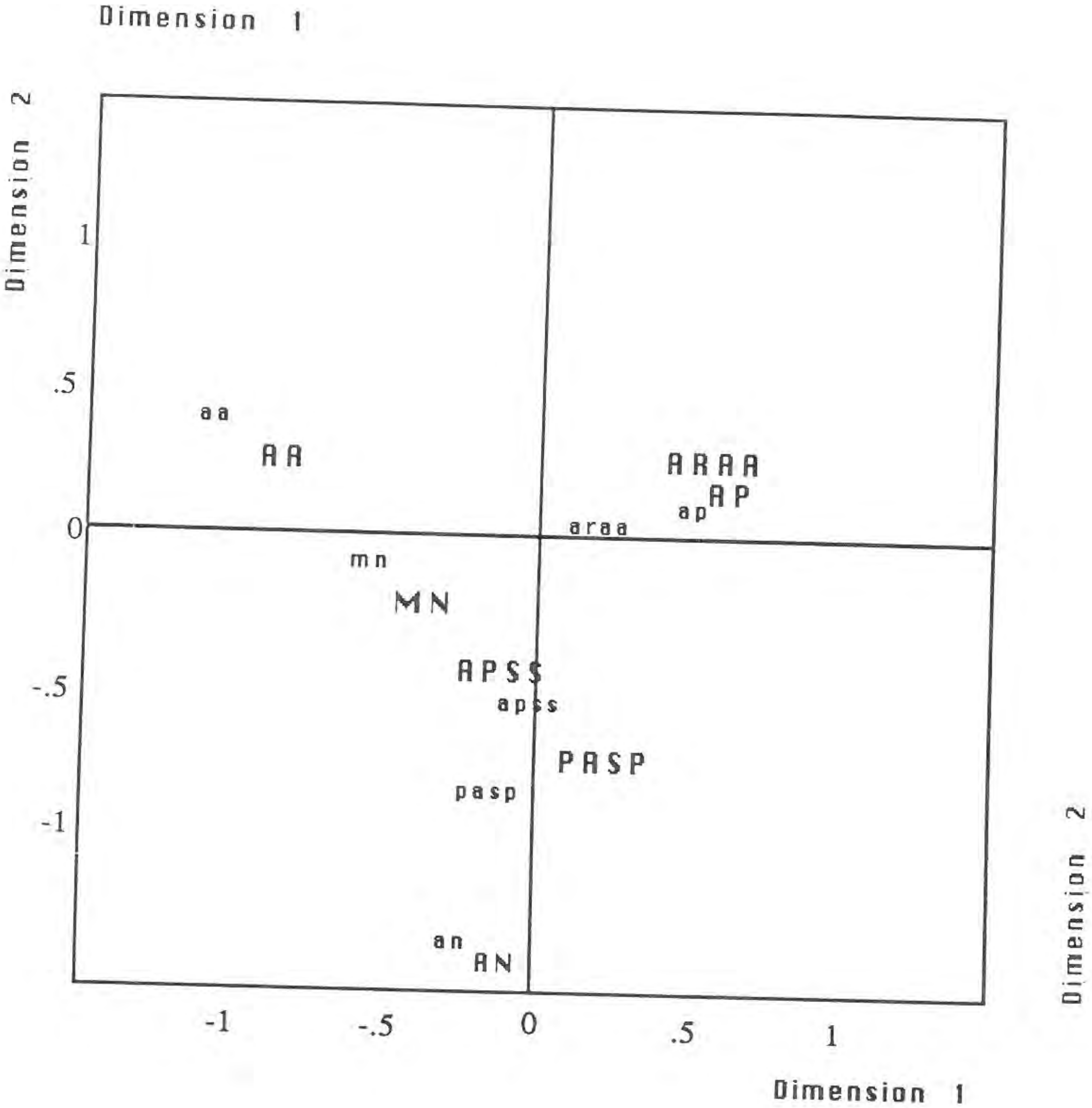
11 J.P. Benzecri et al., *Analyse des donnees* (2 vols.), Paris: Dunod, 1973.

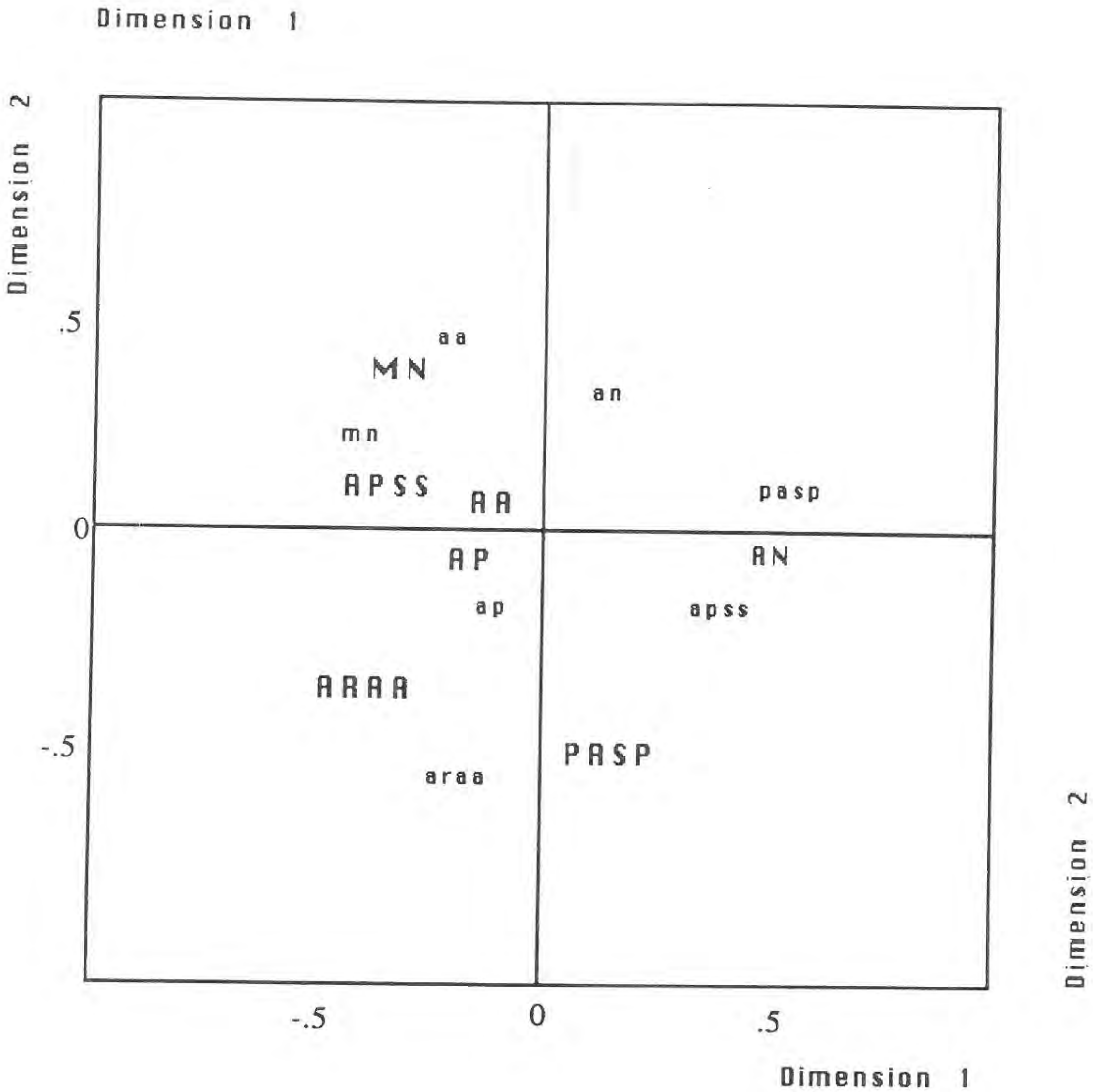
12 P.G.M. van der Heijden, Correspondence analysis of transition matrices, *Kwantitatieve Methoden* 19 (1985) 19-42.

13 J. de Leeuw & P.G.M. van der Heijden, *Quasi-Correspondence analysis*, Research Report RR-85-19, Department of Data Theory, University of Leiden, 1985.

14 E. Garfield (ed.), *SCI Journal Citation Reports*, 18, Philadelphia, PA : Institute for Scientific Information, 1983.

15 A. Gifi, *ANACOR*, Department of Data Theory, University of Leiden, 1985.


Dimension 1
Dimension 2
1
.5
0
-.5
-1
aa
AA
ARAA
ap AP
araa
mn
MN
APSS
apss
PASP
pasp
an
AN
-1
-.5
0
.5
1
Dimension 2
Dimension 1


Dimension 1
Dimension 2
.5
0
-.5
aa
MN
an
mn
APSS
AA
pasp
AP
AN
ap
apss
ARAA
PASP
araa
-.5
0
.5
Dimension 2
Dimension 1