Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 249–255
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Transformations of basic publication–citation matrices

Liming Liang [a,b], Ronald Rousseau [b,c,*]

[a] *Institute for Science Technology and Society, Henan Normal University, Xinxiang 453007, China*
[b] *University of Antwerp (UA), IBW, Universiteitsplein 1, 2600 Wilrijk, Belgium*
[c] *KHBO (Association K.U. Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium*

Received 28 July 2006; received in revised form 22 January 2007; accepted 22 January 2007

Abstract

Basic publication–citation matrices are used to calculate informetric indicators such as journal impact factors or *R*-sequences. Transforming these publication–citation matrices clarifies the construction of other indicators. In this article, some transformations are highlighted together with some of their invariants. Such invariants offer a rigorous mathematically founded way of comparing informetric matrices before and after a transformation.

Keywords: Basic publication–citation matrices; Generalized impact factors; *R*-sequences; Informetric transformations; Invariants

1. Introduction: the basic publication–citation matrix

A basic publication–citation matrix, in short: *p–c* matrix, is a table showing publication and citation data needed for the calculation of an informetric indicator such as a journal impact factor or an *R*-sequence (Liang, 2005). Examples of the use of *p–c* matrices can be found in (Frandsen & Rousseau, 2005; Ingwersen, Larsen, Rousseau, & Russell, 2001; Liang, 2005). Data in such matrices are usually shown in chronological order. Publication data may be given either row by row, as we will do in this article, or column-by-column, as done, e.g. in (Ingwersen et al., 2001): these two approaches are logically equivalent.

Consider Table 1: it shows the number of published articles and citations for a hypothetical set of articles under study. Such an article set is traditionally a journal, i.e. all articles published by a particular journal, but it may also be an institute or a country or any other entity of interest. Recall that citations originate from a pool of citing objects such as all articles published in journals covered by Thomson Scientific. Again, also this pool may vary depending on the type of study. Other alternatives include: all journals covered by one particular JCR category, or all journals covered by a regional citation database (Jin & Wang, 1999; Wu et al., 2004). Recently, a framework for impact calculations and generalizations, taking the pool explicitly into account has been proposed by Rousseau (2005) under the name *conglomerate*.

The first column and the first row of Table 1 refer to periods. In classical citation studies time intervals, denoted as I_k, $k = 1, \ldots, \max(n,m)$, are years, but it is quite feasible to use other periods such as quarters of a year, on the one hand, or decennia on the other. The second column gives the number of published articles by this particular article

* Corresponding author.
E-mail addresses: pllm@public.xxptt.ha.cn (L. Liang), ronald.rousseau@ua.ac.be, ronald.rousseau@khbo.be (R. Rousseau).

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.01.004

Table 1
A basic publication–citation matrix for a hypothetical set of articles

		Interval 1	Interval 2	...	...	Interval n
Interval 1	P_1	$C_{1,1}$	$C_{1,2}$			$C_{1,n}$
Interval 2	P_2		$C_{2,2}$			
⋮					$C_{i,j}$	
⋮						
Interval m	P_m					$C_{m,n}$

set. Concretely, P_i denotes the number of articles published by this article set during period i. The other columns are citation columns. The symbol $C_{i,j}$ denotes the number of citations received in year j by items published in year i. If $i > j$ then $C_{i,j}$ is not defined, or may be set equal to zero (we do not take preprint citations into account). Such zeros are not shown in the publication–citation table. The submatrix of the p–c matrix consisting of all $C_{i,j}$ ($i = 1, \ldots, m; j = 1, \ldots, n$) is called the citation submatrix (in short: the c-submatrix) of the p–c matrix. The first column, providing publication data, is called the publication submatrix (column) or the p-part. Observe also that in general $m \neq n$. The whole matrix consisting of the publication submatrix as its first column, and the citation submatrix is the p–c matrix, denoted as M. To the best of our knowledge such p–c matrices were first described and used by Moed, Burger, Frankfort and van Raan (1983).

In this article, we will propose some useful transformations of this p–c matrix. These transformations will act on the complete p–c matrix M, but may leave some submatrices unchanged. Use of these transformations will be illustrated in the framework of generalized impact factors and of Liang's theory of rhythm indicators (Liang, 2005). In the spirit of Klein's Erlangen program (1872) we will introduce invariants of transformations of the publication–citation matrix. Recall that the German mathematician Felix Klein (1849–1925) proposed – successfully – a new approach to geometry (Boyer and Merzbach, 1989). According to the Erlangen program geometry is the study of invariants of a set under certain group transformations. These groups actually classify the different types of geometries, or stated otherwise: any classification of groups of transformations becomes a codification of geometries. In our case, these invariants tell us which information presented in a p–c matrix stays unchanged through a particular transformation, and which information is altered, the ultimate aim being a classification of publication–citation studies.

2. Generalized impact factors and their relation with rhythm indicators

In Frandsen and Rousseau (2005), a general approach to the notion of an impact factor has been introduced. In this framework, an analysis based on several years of publications becomes possible. In traditional synchronous or diachronous citation studies only one row or one column of the citation submatrix is used. In a synchronous approach the citation period is fixed, in the diachronous approach the publication year (which is the cited year) stays fixed. The more general Frandsen–Rousseau approach to citation studies considers a larger basis than traditional studies. In this type of studies, publication and citation intervals are usually years.

The following notation is used:

n_p denotes the length of the publication window,
n_c denotes the length of the citation window,
Y_p is the first year of the publication period,
Y_c is the first year of the citation period.

Two alternatives are proposed in Frandsen and Rousseau (2005): the so-called general impact factor and the alternative impact factor. The general impact factor of an article set S, denoted $\mathrm{IF}_S(n_p, n_c, Y_p, Y_c)$, is defined as:

$$\mathrm{IF}_S(n_p, n_c, Y_p, Y_c) = \frac{\sum_{i=0}^{n_p-1} \sum_{k=0}^{n_c-1} C_{Y_p+i, Y_c+k}}{\sum_{i=0}^{n_p-1} P_{Y_p+i}} \quad (1)$$

This formulation counts from the first year in the publication period and citation period to the last similar to the calculation of the diachronous impact factor although it also can be used for synchronous analyses. In this notation, the classical Garfield–Sher impact factor of journal J in the year Y is $\mathrm{IF}_J(2, 1, Y-2, Y)$. Indeed:

$$\mathrm{IF}_J(2, 1, Y-2, Y) = \frac{\sum_{i=0}^{1}\sum_{k=0}^{0} C_{Y-2+i,Y+k}}{\sum_{i=0}^{1} P_{Y-2+i}} = \frac{C_{Y-2,Y} + C_{Y-1,Y}}{P_{Y-2} + P_{Y-1}} \tag{2}$$

The 'alternative impact factor' of a set of articles S, denoted $\mathrm{AIF}_S(n_\mathrm{p}, n_\mathrm{c}, Y_\mathrm{p}, Y_\mathrm{c})$, is defined as:

$$\mathrm{AIF}_S(n_\mathrm{p}, n_\mathrm{c}, Y_\mathrm{p}, Y_\mathrm{c}) = \frac{\sum_{i=0}^{n_\mathrm{p}-1}\sum_{k=0}^{n_\mathrm{c}-1} C_{Y_\mathrm{p}+i,\, Y_\mathrm{c}+i+k}}{\sum_{i=0}^{n_\mathrm{p}-1} P_{Y_\mathrm{p}+i}} \tag{3}$$

For example, $\mathrm{AIF}_S(1, 3, Y, Y)$ denotes a 3-year diachronous impact factor. It refers to the publication year Y and uses citations received in three consecutive years, beginning in the publication year Y. The immediacy index of journal J in the year Y is $\mathrm{AIF}_J(1, 1, Y, Y)$, but it may also be written as $\mathrm{IF}_J(1, 1, Y, Y)$. Finally, $\mathrm{AIF}_S(s, 1, Y, Y+t)$ is another interesting generalized impact factor:

$$\mathrm{AIF}_S(s, 1, Y, Y+t) = \frac{\sum_{i=0}^{s-1} C_{Y+i,Y+t+i}}{\sum_{i=0}^{s-1} P_{Y+i}} \tag{4}$$

It denotes the average number of citations per article, over a period of length s, received t years after publication. Here, and in all other examples we assume that the data necessary for the calculations are available in the p–c matrix. As we assume that the reader is familiar with generalized impact factors we do not go into details. Moreover, in this article, impact calculations are just provided as an example of the use of a p–c matrix.

Consider now a fixed $n \times n$ p–c matrix (from now on we take $m = n$). Because of the dimensions of such a table we can calculate $\mathrm{AIF}_S(n-k+1, 1, I_1, I_k)$, $k = 1, \ldots, n$, but not $\mathrm{AIF}_S(s, 1, I_1, I_k)$, $s > n-k+1$.

The creation of the rhythm indicator (Liang, 2005) includes a key measure namely C_k, which denotes the average number of citations per paper in the kth year after its publication ($k = 1$ to $n-i+1$, where $k = 1$ refers to the publication year) and which was constructed as the expected value of the observed $C_{i,i+k-1}$ for any possible i. Just like the general impact factor IF and the alternative impact factor AIF, the rhythm indicator is created from the p–c matrix. In the notation of Table 1, C_k is equal to

$$\frac{\sum_{i=1}^{n-k+1} C_{i,i+k-1}}{\sum_{i=1}^{n-k+1} P_i} \tag{5}$$

The rhythm indicator compares $E_i = P_i \sum_{k=1}^{n-(i-1)} C_k$ with $O_i = \sum_{j=i}^{n} C_{i,j}$ by forming the R-sequence $R_i = O_i/E_i$. It can easily be seen that C_k is just a special AIF, namely: $C_k = \mathrm{AIF}(n-k+1, 1, Y_1, Y_k)$, though the background and original intention for creating C_k and AIF are totally different.

3. The *R*-transformation

This transformation on a p–c matrix M is denoted as R. It maps the matrix M to the p–c matrix $R(M)$. It leaves the publication column P unchanged: the first column of $R(M)$ is equal to P. R transforms the c-submatrix C with elements $C_{i,j}$ into the c-submatrix of $R(M)$ with elements $R_{i,j}$ where

$$R_{i,j} = \frac{P_i}{\sum_{k=1}^{n-j+i} P_k} \sum_{s=1}^{n-j+i} C_{s,s+j-i} = P_i \frac{\sum_{s=1}^{n-j+i} C_{s,s+j-i}}{\sum_{k=1}^{n-j+i} P_k} \tag{6}$$

The first equation establishes the matrix elements $R_{i,j}$ as a weighted sum of citations, the second one as the number of publications multiplied by a generalized impact factor (5).

The P-column is a trivial invariant for the R-transformation. More importantly, the R-transformation leaves the total sum of the elements of the citation submatrix invariant. Hence, the R-transformation can be considered as a rearrangement of the p–c matrix. The proof of this statement is given in the following theorem.

Theorem 1.

$$\sum_{i=1}^{n}\sum_{j=i}^{n} C_{i,j} = \sum_{i=1}^{n}\sum_{j=i}^{n} R_{i,j} \tag{7}$$

Proof

We will show that Eq. (7) is just a reformulation of the rearrangement equation $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$, proved in the appendix of Liang (2005). Indeed, $\sum_{i=1}^{n} O_i$ is by definition equal to $\sum_{i=1}^{n}\sum_{j=1}^{n} C_{i,j}$. Further,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=i}^{n} R_{i,j} &= \sum_{i=1}^{n}\left(\sum_{j=i}^{n}\frac{P_i}{\sum_{t=1}^{n-j+i}P_t}\sum_{s=1}^{n-j+i} C_{s,s+j-i}\right)\\
&= \sum_{i=1}^{n} P_i\left(\sum_{j=i}^{n}\frac{\sum_{s=1}^{n-j+i} C_{s,s+j-i}}{\sum_{t=1}^{n-j+i}P_t}\right)\\
&= \sum_{i=1}^{n} P_i\left(\sum_{j=i}^{n} C_{j-i+1}\right) \qquad (\text{put } k = j-i+1)\\
&= \sum_{i=1}^{n} P_i\left(\sum_{k=1}^{n-i+1} C_k\right)\\
&= \sum_{i=1}^{n} E_i
\end{aligned}$$

This proves Theorem 1. □

Theorem 1 shows that the *R*-transformation is the basic transformation underlying the theory of *R*-sequences as studied in Liang (2005), see also Liang, Rousseau, and Shi (2005) and Liang, Rousseau, and Shi (2006).

4. The AV-transformation

The more (citable) articles a journal, an institute, etc. publishes the larger its citation potential. In order to take this effect into account one may use the following averaging transformation, denoted as AV. Applied to the matrix M this yields the matrix $\mathrm{AV}(M) = A$. The elements of $\mathrm{AV}(M)$ are equal to the elements of M divided by P_i, the first element of the ith row of M:

$$A_{i,j} = \frac{M_{i,j}}{P_i} \tag{8}$$

In particular, we see that the first column of A consists of ones. The citation part of the AV-transformed matrix shows the average number of citations received per published article. The elements of the matrix A are used when calculating diachronous impact factors. The matrix A itself is called the mean p–c matrix.

Deriving elements denoted A_k from $\mathrm{AV}(M)$ in a similar manner as C_k has been derived from M, yields (see Eq. (5)):

$$A_k = \frac{\sum_{i=1}^{n-k+1} A_{i,i+k-1}}{n-k+1} = \frac{1}{n-k+1}\sum_{i=1}^{n-k+1}\frac{C_{i,i+k-1}}{P_i} \tag{9}$$

5. The $\bar{R}$-transformation

The $\bar{R}$-transformation is the multiplicative analogue of the *R*-transformation. It maps the matrix M to the p–c matrix $\bar{R}(M)$. $\bar{R}$ too leaves the publication column P unchanged. Further, $\bar{R}$ transforms the c-submatrix with elements $C_{i,j}$ into

the c-submatrix $\bar{R}(C)$ with elements $\bar{R}_{i,j}$ where

$$\bar{R}_{i,j} = \prod_{k=1}^{n-j+i} \left((C_{k,k+j-i})^{P_i \left(\sum_{s=1}^{n-j+i} P_s \right)^{-1}} \right) \tag{10}$$

The $\bar{R}$-transformation leaves the product of all elements of the p–c matrix elements invariant. Hence, it can be considered as a multiplicative rearrangement of the p–c matrix. The proof of this statement can easily be derived from the analogous statement about R.

Theorem 2.

$$\prod_{i=1}^{n}\prod_{j=i}^{n} C_{i,j} = \prod_{i=1}^{n}\prod_{j=i}^{n} \bar{R}_{i,j} \tag{11}$$

Proof

Consider the p–c matrix M_{L} Its p-column is equal to the p-column of the matrix M, while the elements of its c-submatrix are each equal to the logarithm of the elements of M's c-submatrix. Applying Theorem 1 to M_{L} yields:

$$\sum_{i=1}^{n}\sum_{j=i}^{n} \ln(C_{i,j}) = \sum_{i=1}^{n}\sum_{j=i}^{n} \left(\frac{P_i}{\sum_{s=1}^{n-j+i} P_s} \sum_{k=1}^{n-j+i} \ln(C_{k,k+j-i}) \right) \quad \text{or}$$

$$\sum_{i=1}^{n}\sum_{j=i}^{n} \ln(C_{i,j}) = \sum_{i=1}^{n}\sum_{j=i}^{n} \left(\sum_{k=1}^{n-j+i} \ln \left((C_{k,k+j-i})^{P_i \left(\sum_{s=1}^{n-j+i} P_s \right)^{-1}} \right) \right) \tag{12}$$

Taking exponentials of both sides of Eq. (12) yields Eq. (11).

This proves Theorem 2. □

Combining the AV- and the $\bar{R}$-transformation yields Theorem 3.

Theorem 3. When applying $\bar{R}$ after AV the multiplicative rearrangement equality becomes:

$$\prod_{i=1}^{n} \bar{A}_i = \prod_{i=1}^{n} \bar{B}_i \tag{13}$$

with $\bar{A}_i = \prod_{j=i}^{n} A_{i,j} = \prod_{j=i}^{n} (C_{i,j}/P_i)$, $\bar{B}_i = \left(\prod_{j=1}^{n-i+1} \bar{K}_j \right)$, and where $\bar{K}_i = \left(\prod_{j=1}^{n-i+1} A_{j,j+i-1} \right)^{1/(n-i+1)}$ while Eq. (11) becomes:

$$\prod_{i=1}^{n}\prod_{j=i}^{n} A_{i,j} = \prod_{i=1}^{n}\prod_{j=i}^{n} \left(\prod_{k=1}^{n-j+i} A_{k,j+k-i} \right)^{1/(n-j+i)} \tag{14}$$

Proof

This follows immediately from Theorem 2 by replacing $C_{i,j}$ by $A_{i,j}$ and all P_i by 1.

Note that the multiplicatively transformed $A_{i,j}$ are actually geometric averages. □

The R, $\bar{R}$ and AV-transformations are internal transformations as they are only based on the given p–c matrix. We will now consider another transformation involving an external factor, namely the used pool.

6. Normalizing with respect to the size of the pool

The potential number of retrieved citations clearly depends on the size of the used database (pool). As the p–c matrix usually covers several periods it is often a good idea to normalize it with respect to the size of the pool during that period. Let Z_j denote the size of the pool in interval I_j, $j = 1, \ldots, n$. Then the size-normalized p–c matrix is denoted as N. With respect to the original case the publication column does not change, but $C_{i,j}$ is transformed into $N_{i,j} = C_{i,j}/Z_j$.

This transformation is especially useful when studying relatively small databases or large time periods. Indeed, taking the database SCI-extended (SCIE) as an example, we know that the total number of SCIE's records has been

increasing over the past half century. In previous publications (Liang et al., 2005, 2006) we have shown the influence of this increase on the *R*-sequence of the journal *Science*. Comparing *Science*'s *R*-sequence calculated based on the normalised *p–c* matrix with that calculated based on the original *p–c* matrix we found that for the earlier years values of the *R*-sequence increased after normalization, while for the more recent years they decreased. The same phenomenon occurred for *Science*'s impact factor.

If one wants the total sum of elements in the *c*-submatrix to be invariant for this type of transformation than it must be defined slightly differently, namely as:

$$N'_{i,j} = \frac{C_{i,j}}{Z_j} \frac{\sum_{k=1}^{n}\sum_{s=k}^{n} C_{k,s}}{\sum_{k=1}^{n}\sum_{s=k}^{n}(C_{k,s}/Z_s)} \tag{15}$$

For a given *p–c* matrix and a given pool (of possibly variable size over time) $N'_{i,j}$ is just $N_{i,j}$ multiplied by a normalizing factor.

We note that normalizing with respect to a pool may be combined with an AV-transformation, leaving the details to the reader.

7. Other types of *p–c* matrices

A. Instead of the matrix *C*, where $C_{i,j}$ denotes the total number of citations received in the year *j* by articles published in a particular journal in the year *i*, one may also use the simpler matrix *T*, where $T_{i,j}$ denotes the total number of articles published in this journal in the year *i* and cited in the year *j* (ignoring the precise number of citations each article received, hence $T_{i,j}$ is equal to P_i minus the number of articles that are uncited in the year *j*).
B. Besides a 'cited' perspective, it is also interesting to consider a 'citing behaviour', i.e. a given references perspective. In this case the meaning of the *p–c* matrices changes considerably. Consider again Table 1, now interval I_2 refers to the period before interval I_1. So, if I_1 refers to year *Y*, then I_2 refers to year $Y-1$, and in general I_k refers to year $Y-k+1$. In the citing (or referencing) case the symbol $C_{i,j}$ denotes the number of citations given in articles published in period I_i to articles published in period I_j.

8. *p–c* matrices with discrete steps

In the *p–c* matrices studied thus far rows and columns reflect time periods. Yet, in some studies it is also meaningful to organize the *p–c* matrix by discrete steps, more concretely: each row and column refers to exactly one article or one journal issue, presented in the order in which they are written or published. This approach is especially interesting in self-citation studies, see e.g. (Glänzel, Thijs, & Schlemmer, 2004). We consider the two examples of (1) all articles published by one scientist, and (2) the different issues of one particular journal. In the first example P_i refers to one specific article (hence P_i is always 1) or the number of articles published in journal issue *i*. In the case of articles published by one scientist the symbol $C_{i,j}$ is either zero or one: one if article *i* is cited by article *j* ($j>i$), zero if article *i* is not cited by article *j*. In the case of journal issues the symbol $C_{i,j}$ denotes the number of self-citations issue *i* receives from issue *j* ($j>i$). All transformations and invariants considered for a basic *p–c* matrix can be applied, with a slightly different meaning, to these discrete step *p–c* matrices.

9. Discussion and conclusion

The use of basic publication–citation matrices for the construction of informetric indicators has been highlighted. Transforming these publication–citation matrices clarifies the construction of other indicators. Examples of such transformations, such as the *R* and the AV-transformation are presented. A distinction has been made between transformations using only elements of the given *p–c* matrix, and transformations using external elements, e.g. the size of the citation pool.

In the spirit of Klein's Erlangen program (1872) we introduced invariants of transformations of the publication–citation matrix. We think that informetrics as the mathematical study of information objects may benefit by the introduction of some of the ideas of the Erlangen program. Invariants offer a mathematically founded way of comparing informetric matrices before and after a transformation. For this reason, we think that invariants and the study of invariants should receive more attention in theoretical informetrics.

We are convinced that similar transformations, properties and invariants can be studied for other matrices used in informetric studies, e.g. collaboration matrices or more generally all types of co-occurrence studies.

Acknowledgements

The authors thank Leo Egghe for pointing out that the proof of Theorem 2 can easily be derived form Theorem 1. They also thank the anonymous referees for helpful observations. This work is sponsored by the National Natural Science Foundation of China (Project 70373055).

References

Boyer, C. B., & Merzbach, U. C. (1989). *A history of mathematics*. New York: Wiley.

Frandsen, T. F., & Rousseau, R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology*, *56*, 58–62.

Glänzel, W., Thijs, B., & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, *59*, 63–77.

Ingwersen, P., Larsen, B., Rousseau, R., & Russell, J. M. (2001). The publication–citation matrix and its derived quantities. *Chinese Science Bulletin*, *46*, 524–528.

Jin, B. H., & Wang, B. (1999). Chinese Science Citation Database: Its construction and application. *Scientometrics*, *45*, 325–332.

Liang, L. (2005). The *R*-sequence: A relative indicator for the rhythm of science. *Journal of the American Society for Information Science and Technology*, *56*, 1045–1049.

Liang, L., Rousseau, R., & Shi, F. (2005). The rhythm of science, the rhythm of *Science*. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005* (pp. 398–405). Stockholm: Karolinska University Press.

Liang, L., Rousseau, R., & Shi, F. (2006). The rhythm of science, the rhythm of *Science*. *Scientometrics*, *68*, 535–544.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & van Raan, A. F. J. (1983). *On the measurement of research performance: The use of bibliometric indicators*. State University of Leiden: Report Research Policy Unit.

Rousseau, R. (2005). Conglomerates as a general framework for informetric research. *Information Processing and Management*, *41*, 1360–1368.

Wu, Y., Pan, Y., Zhang, Y., Ma, Z., Pang, J., Guo, H., et al. (2004). China Scientific and Technical Papers and Citations (CSTPC): History, impact and outlook. *Scientometrics*, *60*, 385–394.