Averaging and globalising quotients of informetric and scientometric data

Leo Egghe

Limburgs Universitair Centrum, Diepenbeek, and UIA, Wilrijk, Belgium

Ronald Rousseau

UIA, Wilrijk, and Limburgs Universitair Centrum, Diepenbeek, Belgium

Received 26 August 1995
Revised 27 October 1995

Abstract.

Based on the particular case of the average impact factor of a subfield versus the impact factor of this subfield as a whole, the difference is studied between an average of quotients, denoted as AQ, and a global average, obtained as a quotient of averages, and denoted as GQ. In the case of impact factors, AQ becomes the average impact factor of a field, and GQ becomes its global impact factor.

Many applications in the context of informetrics and scientometrics are given, e.g. the Price index, the text to reference ratio, ageing, the receptivity factor for foreign literature, journal price calculation, discipline influence scores, and fill-rates as measures of library performance. We strongly claim that, in most applications, the global average is the preferred one.

It is also shown that, if geometric averages are used instead of arithmetic ones in the definition of AQ and GQ, the difference between the two approaches is eliminated.

Correspondence to: Professor Dr L. Egghe, Limburgs Universitair Centrum, Universitaire Campus, B–3590, Diepenbeek, Belgium. Tel: +32 011 26 81 21. Fax: +32 011 26 81 26. E-mail: legghe@luc.ac.be

1. Introduction

It is not difficult, using ISI's *Journal Citation Reports* (*JCR*), to calculate average impact factors for *JCR*'s subject categories. Yet, it is certainly more interesting, e.g. for science evaluation purposes, to know the global impact factor of a subject category [1]. Van Hooydonk *et al.* [2, 3] and Egghe and Rousseau [4] have studied differences between the average impact factor (AIF) and the global impact factor (GIF). The most interesting theoretical result is that the slope of the regression line of the impact as a function of the number of publications is positive if, and only if, the global impact is larger than the average impact. As a corollary, we found that if the impact considered as a function of the number of publications is increasing then the GIF is larger than the AIF. In this contribution, we will study, in general, similar relations to that between the GIF and the AIF. We will show that here also interesting and, in fact, similar results can be proved. Next, we will demonstrate the relevance of these considerations for different notions in informetrics and scientometrics, such as the Price index, text to reference ratio, receptivity factor, journal prices, discipline influence scores, fill-rates as measures of library performance, ageing and, finally, gross regional product (an econometric application). Most mathematical calculations can be found in the Appendices.

2. A general framework

Consider a set of N entities, e.g. journals, of which we measure or count two properties, e.g. the number of articles published in each journal during a fixed period and the number of citations to these articles over a certain period. The obtained numbers are denoted $(x_i)_{i=1,\ldots,N}$ and $(y_i)_{i=1,\ldots,N}$. We will always

assume that the x_i are strictly positive. We are interested in the quotients $(Q_i)_{i=1,\ldots,N} = (y_i/x_i)_{i=1,\ldots,N}$ for which there usually exists a distinct name; e.g. for the case of journals, the quotient 'citations per publication' is known as the journal's impact or impact factor. Now, we want to find a number describing an average 'quotient property' for the whole set of entities. Two obvious approaches immediately spring to mind: either one can take the average of all quotients Q_i, which we will denote by AQ, or one can sum all the y_i and divide this number by the sum of all the x_i. This will be denoted as GQ (global Q). Formally:

$$AQ = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{X_i}$$

and

$$GQ = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

where μ_y is the average of the y-values, and μ_x is the average of the x-values. The ratio GQ/AQ will be denoted by ρ_Q.

In practical situations, it often happens that data are obtained one by one, and that it takes a long time before all data are known. In those situations, it is desirable to compute AQ and GQ values, based on partial data, and to be able to update them in a simple manner. Appendix 1 provides such a way, using a recursion relation. Further, we note that the variable GQ can be described as a weighted sum of the Q_i (see Appendix 1).

3. Mathematical results

For every i, i = 1, . . . , N we have an x_i and a y_i value. Hence, with every x_i there corresponds a y_i. This means we have a relation from the set of x-values to the set of y-values. Hence, the ys are related to the xs. Consequently, also the Q-values, being quotients of y- and x-values, can be related to the x_i. Now, we can compute the regression line of Q over x. The slope of this regression line is denoted as r_x. This leads to the following relation: $r_x > 0$ if, and only if, $\rho_Q > 1$, i.e. this regression line is increasing if, and only if, ρ_Q is larger than 1, i.e. GQ > AQ. Moreover, the equality $r_x = 0$ occurs if, and only if, $\rho_Q = 1$. The proof, based on the mathematical expression for the slope of a regression line (see, e.g. [5, p. 66]) can be found in Appendix 2.

As a corollary, we note that if Q is an increasing function of x, then $\rho_Q > 1$; similarly, if Q is a decreasing function of x, then $\rho_Q < 1$. For a formal proof, refer to [4, Appendix]. For the case of an impact factor, this means that if the impact factor of a journal generally increases with the number of published articles (a fact observed and described in [3]), then the global impact is larger than the average one.

Finally, we note that we obtained a decomposition of the global value (GQ) into the average one (AQ) and a term depending on the weight distribution of the x_is over the Q_is. Also, this result is shown in Appendix 2.

4. Practical examples

Before presenting practical cases where GQ and AQ are used, we want to emphasise the fact that, when studying a global property calculation of the global value, GQ is highly recommended.

4.1. AIF and GIF

As mentioned in Section 1, the AIF and the GIF of a scientific subfield [2, 3, 4] provided the motivating example for studying the relations of Section 3.

4.2. Price index

For an article, one counts the total number of references and the number of references which are given to articles not older than d years, where the year of publication is counted as year one. (Price [6] uses d = 5.) Its quotient is Price's index. We observe an ambiguity here in the existing literature: some authors, such as Price, state that they use the first five years. In this terminology, it is unclear whether the year of publication is the year zero or the year one. Moreover, it is unclear whether or not this year zero is included. Moed [7] uses the average Price index as a measure for the field, while Price himself used the global one. Of course, one can equally well calculate Price's index for a journal. The difference between the two approaches was discussed by Wouters and Leydesdorff [8]. They note that for the journal *Scientometrics* the average – five-year – Price index was 0.514, while the global one was 0.43. From the fact that the global one is smaller than the average one, we conclude that the slope of the regression line of Price's index over the total number of references is decreasing. Price's index was further studied in [9].

4.3. Text to reference ratio

In [10], a number of journals was randomly selected and, for each article published during the years 1980 and 1987, the (estimated) number of words and the number of references were obtained. Its quotient yields the text to reference ratio of each article. As far as we could see, the authors did not state how they obtained the journal data; namely, as an average of article data or as global text to reference ratios. We assume they used the global method. In this investigation, the authors considered one more level of aggregation. They grouped the articles according to field and calculated text to reference ratios for each field (botany, physical chemistry, geology).

4.4. Receptivity factor

In this application, the field is fixed. For every article under consideration, one collects the number of references and the number of references to articles written by fellow countrymen. This is done per country, where, again, it is possible to take an average or a global point of view. Dividing this result by the share of the country in the total output in the field yields the receptivity index for foreign literature. Herman [11] has calculated this index for the UK and USA in the field of library and information science. Note that the division per country was not done according to the nationality of the author but according to the nationality of the journal. Calculating receptivity factors for different countries and different fields is one of the suggestions for research in publication and citation analysis listed by Rousseau [12].

4.5. Journal prices

For every journal, one collects the number of published pages (in a year) and the subscription price. The quotient is the price per page. One could similarly calculate the price per character, or the price per citation, as a kind of 'value for money' indicator. Results can then be brought together per field or per publisher [13]. For some publishers, this seems to be a very controversial procedure [14, 15]!

Van Hooydonk *et al.* [2] calculated a price per article. The authors give preference to the weighted, i.e. global, price per article of a discipline or faculty, as they also do for the case of the impact factor.

4.6. Discipline influence score

For a fixed journal A and a fixed set of journals, representing a discipline, one collects for every journal in the set the total number of times this journal cited all other journals of the discipline (during a fixed period) and the number of times the journal cites journal A. Then the quotients Q_i are formed:

$$Q_i = \frac{\text{number of times journal i cited journal A}}{\text{total number of times journal i cited all journals}}$$

The average value of the Q_is (AQ) is what Hirst [16] defines as the discipline influence score of journal A. As far as we know, the corresponding GQ has not been used, although we would prefer this latter index. Hirst's discipline influence score was also the basis for the discipline-specific journal selection algorithm proposed by He and Pao [17].

4.7. Fill-rates as measures of library performance

Materials fill-rates of libraries refer to the ability of a user to locate library materials. As pointed out by D'Elia [18] and discussed by Van House [19], such fill-rates are not unambiguous measures of library performance: they measure as well the ability of the library patron to conduct successful searches. As individuals differ in their abilities and in the number of searches performed, D'Elia suggests that one might calculate individual fill-rates (these are the Qs of our general theory, the xs being the number of performed searches, and the ys the number of successful searches) and then to take the average of these individual fill-rates rather than taking global fill-rates. In other words, he prefers AQ above GQ. We do agree with D'Elia's remark that a patron's ability is not entirely within the library's control, and hence fill-rate in this sense is not an unbiased measure of library performance, yet we cannot see why the average fill-rate is a better measure than the global one. Van House [19] prefers the global approach, mainly for practical reasons.

4.8. Ageing

For each article, one counts c_j, the number of references to articles which are j years old, $j = 0, 1, \ldots, 10$. Here, the number 10 is used for convenience; we further assume that none of the c_j is equal to zero. Then the ageing rate r_k of this article is determined as the average of the quotients c_{j+1}/c_j, $j = 0, \ldots, 9$. Note that other definitions of ageing are used in the literature. The above definition is used only as an example.

Then, to determine the ageing rate of a journal, one can use the average of the ageing rate of all articles (AAR: average ageing rate of a journal), or one can use a global approach, i.e. take the sum of all c_js, form quotients and then take the average of all quotients (GAR: global ageing rate of a journal). Assuming that there are N articles in this journal, we have:

$$AAR = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{10}\sum_{j=0}^{9}\frac{C_{j+1}^{(k)}}{C_j^{(k)}}\right)$$

$$GAR = \frac{1}{10}\left(\sum_{j=0}^{9}\frac{\sum_{k=1}^{N}C_{j+1}^{(k)}}{\sum_{k=1}^{N}C_j^{(k)}}\right)$$

This is an example of a multi-layered approach.

4.9. Gross regional product (GRP)

Averages, as discussed in this article, occur in many fields. A well-known example from econometrics (recall the appeal for more interdisciplinarity in the information sciences; in particular, in relation to econometrics [20, 21]) is the case that for every country one collects the gross national product (GNP) and the number of inhabitants. Its quotient yields the GNP per capita. When considering, for example, the GRP per capita in the European Union, one can take the average of all GNP per capita of every member country, or one could calculate the global GRP per capita. Again, we think that the second index is the more significant one. By the way, it is well known (see, e.g. [22]) that there exists a direct – though not linear – relationship between the scientific production of a country and the GNP per capita. Recall further that the GNP has been used by Rousseau and Vervliet [23] as a predictor for the potential interlending demand in the European Union.

5. The geometric mean

We strongly endorse the use of global measures, unless there are special reasons to use the average. In this case, it must be clearly stated that the 'global' result is obtained by averaging over sub-populations.

One could wonder if there is no way to eliminate this difference between global and average measures. Although we think that this difference is useful and, in general, should not be eliminated, we are able to give another measure which does not make this difference. It suffices to use the geometric mean. Indeed, given the sequences $(y_i)_{i=1,\ldots,N}$ and $(x_i)_{i=1,\ldots,N}$ we can form the Q-sequence $(Q_i)_{i=1,\ldots,N} = (y_i/x_i)_{i=1,\ldots,N}$ and consider then its geometric mean, denoted as A_GQ:

$$A_GQ = \left(\frac{y_1}{x_1}\frac{y_2}{x_2}\ldots\frac{y_N}{x_N}\right)^{\frac{1}{N}}$$

Now:

$$GQ = \frac{\mu_y}{\mu_x}$$

so that it is natural to define its geometric analogue, denoted as G_GQ, as:

$$G_GQ = \frac{(y_1 \ldots y_N)^{\frac{1}{N}}}{(x_1 \ldots x_N)^{\frac{1}{N}}}$$

which is clearly equal to A_GQ.

Note: Starting from data collected on small sets, calculating a property on some aggregation level and then going to a higher aggregation level (and possibly further) can be done in other ways than by calculation quotients and averages. A case in point is the calculation of concentration. Here also, one prefers measures such as Theil's, which can be decomposed into a term coming from the sublevels and a term coming from the globalisation (see, e.g. [24, p. III.5]).

6. Conclusion

We have studied the measures AQ (average of quotients) and GQ (global average, or quotient of averages), where:

$$AQ = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{x_i} \text{ and } GQ = \frac{\sum_{i=1}^{N}y_i}{\sum_{i=1}^{N}x_i} = \frac{\mu_y}{\mu_x}$$

It is shown that GQ > AQ if, and only if, the slope of the regression line of the scatterplot $(x_i, y_i/x_i)_{i=1,\ldots,N}$ is positive. We further presented a decomposition of GQ into AQ and a term depending on the weight distribution of the x_is over the Q_is (= y_i/x_i). Next, we have illustrated the relevance of these considerations

for different fields of informetrics, such as the Price index, text to reference ratio, receptivity factor, journal prices, discipline influence scores, fill-rates as measures of library performance, ageing and, finally, GRP (an econometric application). We strongly claim that, in most applications, the global average is the preferred one of the two measures. Yet, if one wishes to eliminate the differences between the two approaches, geometric averages can be used instead of arithmetic ones in the definition of AQ and GQ.

Appendices

Appendix 1: The variable GQ described as a weighted sum of the Q_i and a recursion relation for AQ and GQ.

The variable GQ can be described as a weighted sum of the Q_i: indeed,

$$GQ = \frac{\sum_{i=1}^{N} y_i}{\sum_{j=1}^{N} x_j} = \sum_{i=1}^{N} \frac{y_i}{\sum_{j=1}^{N} x_j} = \sum_{i=1}^{N} \frac{x_i}{\sum_{j=1}^{N} x_j} \frac{y_i}{x_i} = \sum_{i=1}^{N} w_i Q_i$$

where the weights w_i, i = 1, . . . N are given as:

$$w_i = \frac{x_i}{\sum_{j=1}^{N} x_j}$$

Both AQ and GQ can be calculated recursively, as follows:

$$AQ\ (N) = \frac{N-1}{N} AQ\ (N-1) + \frac{Q_N}{N}$$

$$GQ\ (N) = (1 - w_N)\ GQ\ (N-1) + w_N Q_N$$

where GQ(N–1) and AQ(N–1) denote GQ and AQ calculated over the first N–1 data points.

Appendix 2: A proof of the relation between ρ_Q and the slope of the regression line (see also [4]) and a decomposition theorem. (For the notation, refer to the main text.)

Theorem

$$r_x > 0 \Leftrightarrow \rho_Q > 1$$

and also:

$$r_X = 0 \Leftrightarrow \rho_Q = 1$$

Proof

$$\rho_Q > 1$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{y_i}{x_i} < N \frac{\mu_y}{\mu_x}$$

On the other hand, the slope r_x of the regression line of Q over x is larger than 0

$$\Leftrightarrow N \sum_{i=1}^{N} x_i Q_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} Q_i\right) > O$$

(see, e.g. [5, p. 66])

$$\Leftrightarrow \sum_{i=1}^{N} \frac{y_i}{x_i} < N \frac{\mu_y}{\mu_x}$$

The result about the equality sign follows similarly. This proves this theorem.

A decomposition result

We show how the global average can be decomposed into the average one and a term depending on the weight distribution of the x_is over the Q_is.

Proposition

$$GQ = AQ - \frac{1}{N} \sum_{i=1}^{N} Q_i \frac{\mu_x - x_i}{\mu_x}$$

Proof

$$GQ = \frac{\mu_y}{\mu_x} = \sum_{i=1}^{N} \frac{y_i}{N \mu_x}$$

$$= \sum_{i=1}^{N} \left[\frac{Q_i - Q_i}{N} + \frac{1}{N} \left(\frac{y_i x_i}{x_i \mu_x} \right) \right]$$

$$= AQ - \frac{1}{N}\sum_{i=1}^{N} Q_i \left(\frac{\mu_x - x_i}{\mu_x}\right)$$

References

[1] R. Rousseau, *A Scientometric Study of the Scientific Publications of LUC (Limburgs Universitair Centrum) Period 1981–1993* (Report, 1995).

[2] G. Van Hooydonk, R. Gevaert, G. Milis-Proost, H. Van de Sompel and K. Debackere, A bibliotheconomic analysis of the impact factors of scientific disciplines, *Scientometrics* 30 (1994) 65–81.

[3] R. Rousseau and G. Van Hooydonk, Journal production and journal impact factors, *Journal of the American Society for Information Science* (1996) [to appear].

[4] L. Egghe and R. Rousseau, *Average and Global Impact of a Set of Journals* [submitted to *Scientometrics*].

[5] L. Egghe and R. Rousseau, *Introduction to Informetrics* (Elsevier, Amsterdam, 1990).

[6] D. De Solla Price, Citation measures of hard science, soft science, technology, and nonscience. In: C.E. Nelson and D.K. Pollack (eds), *Communication Among Scientists and Engineers* (Heath, Lexington, MA, 1970) pp. 3–22.

[7] H.D. Moed, Bibliometric measurement of research performance and Price's theory of differences among the sciences, *Scientometrics* 15 (1989) 473–483.

[8] P. Wouters and L. Leydesdorff, Has Price's dream come true: is scientometrics a hard science? *Scientometrics* 31 (1994) 193–222.

[9] L. Egghe, *The Price Index and Its Relation to the Mean and Median Reference Age* (1995) [preprint].

[10] A.E. Little, R.M. Harris and P.T. Nicholls, Text to reference ratios in scientific journals. In: L. Egghe and R. Rousseau (eds), *Informetrics 89/90* (Elsevier, Amsterdam, 1990) pp. 211–216.

[11] I.L. Herman, Receptivity to foreign literature: a comparison of UK and US citing behavior in librarianship and information science, *Library and Information Science Research* 13 (1991) 37–47.

[12] R. Rousseau, Suggestions for research topics in citation and publication analysis, *Library Science with a Slant to Documentation and Information Studies* 32 (1995) 3–12.

[13] H.H. Barschall and J.R. Arrington, Cost of physics journals: a survey, *Bulletin of the American Physical Society* 33 (1988) 1437–1447.

[14] C. Holden, Gordon and Breach impanels a journal jury, *Science* 249 (1990) 298–299.

[15] A.L. O'Neill, The Gordon and Breach litigation: a chronology and summary, *Library Resources and Technical Services* 37 (1993) 127–133.

[16] G. Hirst, Discipline impact factors: a method for determining core journal lists, *Journal of the American Society for Information Science* 29 (1978) 171–172.

[17] C. He and M.L. Pao, A discipline-specific journal selection algorithm, *Information Processing and Management* 22 (1986) 405–416.

[18] G. D'Elia, Materials availability fill rates – useful measures of library performance? *Public Libraries* 24 (1985) 106–111.

[19] N. Van House, Public library effectiveness: theory, measures, and determinants, *Library and Information Science Research* 8 (1986) 261–283.

[20] L. Egghe, Bridging the gaps: conceptual discussions on informetrics, *Scientometrics* 30 (1994) 35–47.

[21] R. Rousseau, Similarities between informetrics and econometrics, *Scientometrics* 30 (1994) 385–387.

[22] J.D. Frame, National economic resources and the production of research in lesser developed countries, *Social Studies of Science* 9 (1979) 233–246.

[23] R. Rousseau and H.D.L. Vervliet, A prediction of the potential interlending demand in the European Community, *Libri* 40 (1990) 278–294.

[24] R. Rousseau, *Concentration and Diversity in Informetric Research* (PhD Thesis, Antwerp University, 1992).