

Paper presented at the Second International Seminar on Quantitative Evaluation of Research Performance – Shanghai, 23-25 October 2000.

## **Concentration and evenness measures as macro-level scientometric indicators**

---

**Ronald Rousseau**

KHBO – Department of Industrial Sciences and Technology  
Zeedijk 101, B-8400 Oostende, Belgium  
e-mail: ronald.rousseau@kh.khbo.be

### **Abstract**

There exist different types of concentration and evenness measures. What they have in common is the fact that they all try to measure inequality. We will discuss only concentration measures that describe the scatter of items over sources, and this irrespective of the number of sources considered. Concentration measures can, e.g., be used as indicators for the convergence, or divergence of regions. Use of different types of concentration measures as scientometric indicators is illustrated for the publication output of China's different administrative units (provinces, autonomous regions and municipalities).

Three aspects of concentration, each leading to a different Lorenz curve, and resulting in a different partial order, are studied. In each case, we suggest measures respecting the corresponding partial order. Concretely, we study:

- classical Lorenz curves and concentration measures
- comparison with an internal or external standard
- source per source comparison of items

Different forms of concentration and corresponding measures are important indicators for the dynamics of an R&D system. We advance the proposition that this aspect has been neglected too much in the past.

## Introduction

During the Beijing International Seminar of Quantitative Evaluation of R&D in Universities and Fifth All-China Annual Meeting for Scientometrics and Informetrics in December 1998, I had the honour to present a talk on research evaluation at universities and scientific institutes (Rousseau, 2000a). The subject discussed here is more general in its applications, but nevertheless related to the construction of scientometric indicators. We will deal with the study of concentration and evenness and will show how this is related to many important issues such as the study of biodiversity, but also economic inequality of regions, and research evaluation.

Consider, as an introduction, the following true or false questions:

- Science is characterized by a large inequality among the participants in the system: scientists as authors, journals as sources for articles, articles as sources for citations, institutes and countries as producers of scientific knowledge.
- All search engines cover about the same portion of the Internet.
- Incomes in the different regions of China are converging.
- The world income inequality increases, i.e. the gap between the rich and the poor becomes worldwide larger and larger.
- Biodiversity in tropical regions decreases at alarming rates.

Questions like these can only be answered if one has a good theory, and consequently appropriate measures for inequality or concentration. The study of inequality is, however, not an easy subject. There exist different types of concentration measures. What they have in common is the fact that they all try to measure inequality. Yet, there is not a single mathematical definition of concentration. One could say that this notion is only loosely defined. But we find 'loose' definitions in many fields. Even the most used statistical parameter, namely the average, can be defined in several ways: we have the arithmetic average, the geometric average, the harmonic average etc. In this article we explain the difference between several notions of concentration.

Basically, concentration can be described as

*the relative apportionment of items among the sources present.*

We will discuss only concentration measures that describe the scatter of items over sources, and this irrespective of the number of sources considered. The opposite of this kind of concentration is known as 'evenness'. Evenness plays a key role in ecological studies (Nijssen et al., 1998; Rousseau & Van Hecke, 1999). We will phrase most of our findings in terms of concentration but they could equally well be phrased in term of evenness. Measures that also take the number of sources into account were referred to as concentration measures of the second type in (Rousseau, 1992a). In ecology their opposites

are known as diversity or heterogeneity measures (Rousseau & Van Hecke, 1999). These will not be discussed here.

Concentration measures are not quality indicators, at least not directly. If, however, science policy measures are aimed at reducing inequality between, e.g. countries, regions or institutes then these measures can be used as indicators for the success of the program. Indeed, such measures can, among other things, be indicators for convergence or divergence between the entities of the system under study. They may, in general, be used for comparing individuals, research groups, institutions, regions, countries, disciplines etc.

In this article we start from the framework of cross-classified data (pertaining, e.g. to two classifications): a row classification consisting of  $M$  rows, and a column classification consisting of  $N$  columns, with  $M, N \geq 2$ . An analysis of this matrix leads us to different forms of Lorenz curves. These form the basis of partial orders that, finally, lead to concentration measures. Use of different types of concentration measures as scientometric indicators is illustrated for the case of the publication output of China's administrative units, based on the CSCD (Jin & Wang, 1999).

This article is an extended version of a poster presented at the Sixth International Conference on Science and Technology Indicators (Leiden, 24-27 May 2000 (Rousseau, 2000b)).

## Method

We assume that an  $(M \times N)$  matrix, i.e. a two-dimensional table, as in Table A is given.

Table A

An illustrative example. The two-way classification used in this table classifies published articles per journal (J) and per country of the first author (C).

| Countries | C1  | C2  | C3  | C4  | Total:J |
|-----------|-----|-----|-----|-----|---------|
| Journals  |     |     |     |     |         |
| J1        | 10  | 20  | 100 | 70  | 200     |
| J2        | 40  | 60  | 60  | 340 | 500     |
| J3        | 50  | 20  | 40  | 190 | 300     |
| Total: C  | 100 | 100 | 200 | 600 | 1000    |

Based on this table of raw data we derive other data and tables. A vector describing the distribution of articles per country is shown in the last row (Total):  $C = (100,100,200,600)$ . Similarly, a vector describing the distribution of articles per journal can be found in the last column of Table A:  $J = (200,500,300)$ .

Per country, this is per column, one can derive a distribution vector. For C1 this is the vector: (10/100,40/100,50/100). Doing this for each column leads to the following table (Table B) of distribution (column) vectors.

Table B

|    | C1  | C2  | C3  | C4   |
|----|-----|-----|-----|------|
| J1 | 0.1 | 0.2 | 0.5 | 0.12 |
| J2 | 0.4 | 0.6 | 0.3 | 0.57 |
| J3 | 0.5 | 0.2 | 0.2 | 0.32 |

In a similar way one derives distribution vectors per journal (i.e. per row). For J1 this is: (10/200, 20/200,100/200, 70/200). This leads to the following table (Table C) of distribution (row) vectors.

Table C

|    | C1   | C2   | C3   | C4   |
|----|------|------|------|------|
| J1 | 0.05 | 0.10 | 0.50 | 0.35 |
| J2 | 0.08 | 0.12 | 0.12 | 0.68 |
| J3 | 0.17 | 0.07 | 0.13 | 0.63 |

Finally, one can consider relative values for Table A as a whole. By this we mean that we consider the contribution of each cell to the whole matrix. Then the sum of all cells is one. This procedure, applied to Table A leads to Table D.

Table D

|                 | C1   | C2   | C3   | C4   | Relative totals |
|-----------------|------|------|------|------|-----------------|
| J1              | 0.01 | 0.02 | 0.10 | 0.07 | 0.20            |
| J2              | 0.04 | 0.06 | 0.06 | 0.34 | 0.50            |
| J3              | 0.05 | 0.02 | 0.04 | 0.19 | 0.30            |
| Relative totals | 0.10 | 0.10 | 0.20 | 0.60 | 1.00            |

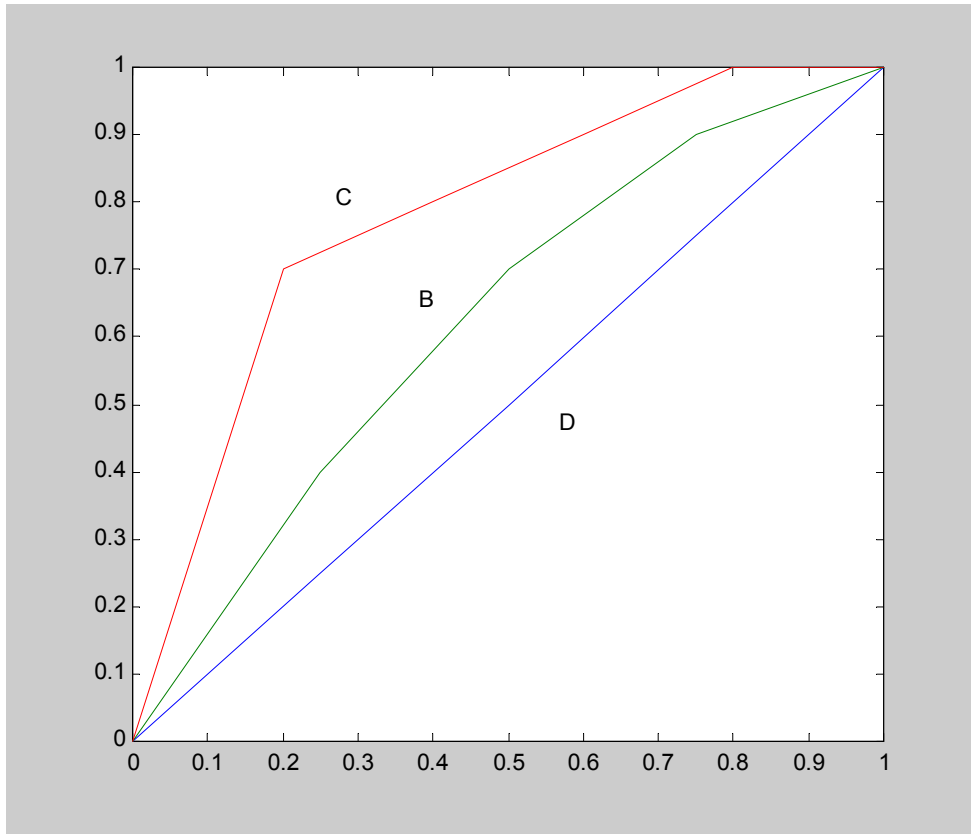
Note that the value of cell (i,j) in Table D is not equal to the product of the corresponding values in Tables B and C. This would only be the case if the variables studied here were statistically independent.

Now, we can study, for each country separately, the concentration of articles over journals. This means that we associate a concentration measure to each column vector in Table B. Similarly, we can study the concentration, for each journal separately, of articles over countries. Then we associate a concentration measure to each row vector of Table C. Concentration in this

sense is represented by classical Lorenz curves (Lorenz, 1905), discussed in the next section.

### **Classical Lorenz curves and concentration measures**

Assume that there are  $N$  sources and let  $X = (x_1, x_2, \dots, x_N)$  be a vector of abundances, i.e.  $x_i$  denotes the number of items 'produced' by the  $i$ -th source,  $i = 1, \dots, N$ . We now recall how classical Lorenz curves are constructed. First, sources are ranked according to their abundances (from highest to lowest). Then cumulative proportions of sources as abscissas are drawn against correspondingly ranked cumulative proportions of items as ordinates. Note the use of proportions:  $N$ -tuples that differ only by a proportionality factor lead to the same Lorenz curve. Similarly, vectors that only differ in the order of their components are represented by the same Lorenz curve. Such vectors are said to be equivalent. Generally, vectors (with a possibly different number of sources) yielding the same Lorenz curve are said to be equivalent. Hence vectors such as  $(1, 2, 3, 4)$ ,  $(2, 4, 3, 1)$ ,  $(1, 1, 2, 2, 3, 3, 4, 4)$  and  $(9, 12, 6, 3)$  all are equivalent. In Fig.1 three lines are drawn. Line D the straight diagonal line, is the Lorenz curve of perfect evenness, i.e. of lowest concentration. This line plays the role of a reference line. If any change occurs then the Lorenz curve is situated above the line of perfect evenness. Lines B and C represent the Lorenz curves of  $(4, 3, 2, 1)$  and  $(7, 1, 1, 1, 0)$ . Lorenz curves determine a partial order, denoted as  $\prec\prec$  in the set of all vectors. Indeed, if the Lorenz curve of a vector  $X$  lies under the Lorenz curve of vector  $Y$  (more precisely, if no part of the Lorenz curve of  $X$  lies strictly above the Lorenz curve of  $Y$ ) then  $X \prec\prec Y$ . Vectors that correspond to the line of perfect evenness are minimum vectors in this poset (partially ordered set). The ordering is only partial in the sense that vectors corresponding to Lorenz curves that intersect are non-comparable.



**Fig.1 “Normal” Lorenz curves of (4,3,2,1) (curve B),  
and (7,1,1,1,0) (curve C),  
D denotes the diagonal ( i.e. the line of perfect evenness)**

Functions that respect this partial order are called concentration measures. Examples are: the modified Simpson (or Herfindahl) index ( $\lambda_m$ ), the coefficient of variation ( $V$ ), the entropy measure ( $H$ ), the Gini index ( $G$ ) and the length of the Lorenz curve (LOR). These are defined as follows (Nijssen et al., 1998).

The modified Simpson or Herfindahl index of a vector  $X$ , denoted as  $\lambda_m(X)$ , is defined as:

$$\lambda_m(X) = N \sum_{j=1}^N a_j^2, \text{ where } a_i = \frac{x_i}{\sum_{j=1}^N x_j} = \frac{x_i}{T} \quad (1)$$

The coefficient of variation is defined as

$$V(X) = \sigma/\mu, \quad (2)$$

where  $\sigma$  denotes the standard deviation and  $\mu$  denotes the mean of the vector  $X$ .

The entropy concentration measure  $H$  is defined as:

$$H(X) = \ln(N) + \sum_{i=1}^N \left(\frac{x_i}{T}\right) \ln\left(\frac{x_i}{T}\right) = \ln(N) + \sum_{i=1}^N a_i \ln(a_i) \quad (3)$$

The Gini index  $G$ : this concentration measure is defined as:

$$G(X) = \frac{N+1}{N} - \frac{2}{N} \sum_{i=1}^N i a_i = \frac{N+1}{N} - \frac{2}{\mu N^2} \sum_{i=1}^N i x_i \quad (4)$$

where the  $x_i$ 's are ranked from high to low and  $\mu$  denotes the mean of the set  $\{x_i\}$ . It can easily be seen that  $G(X)$  is equal to twice the area between the Lorenz curve and the diagonal of perfect evenness. Consequently, the value of the Gini index for the equality situation is equal to zero. As an illustration we mention that because of the (unequal) economic development in China the Gini index of household incomes has increased from 0.382 in 1988 to 0.452 in 1995 (Khan & Riskin, 1998). This answers one of the questions asked in the introduction.

Finally, the length of the Lorenz curve of the vector  $X$  is (Dagum, 1980; Rousseau, 1992a):

$$LOR(X) = \sum_{i=1}^N \sqrt{a_i^2 + \frac{1}{N^2}} \quad (5)$$

Of course one can also study the total concentration of Table A, or equivalently, Table D. This is probably not very interesting, unless one can decompose total concentration according to the contribution of different countries (or journals). The entropy measure is best suited for this purpose. Indeed, if we denote the entropy measure for a vector  $X$  with  $N$  sources as  $H(X, N)$ , and if these sources can be subdivided into  $d$  subpopulations, with  $n_g$  sources each, then we denote the concentration vector of subpopulation  $g$  as  $X^{(g)} = (x_1^{(g)}, \dots, x_{n_g}^{(g)})$ . Its mean, denoted as  $\mu_g$ , is obtained as:

$$\mu_g = \frac{\sum_{i=1}^{n_g} x_i^{(g)}}{n_g} \quad (6)$$

Then, it is easy to show (Theil, 1967; Shorrocks, 1980; Rousseau, 1992a) that

$$H(X, N) = \sum_g \frac{n_g \mu_g}{N \mu} H(X^{(g)}, n_g) + \frac{1}{N} \sum_g n_g \frac{\mu_g}{\mu} \ln \left( \frac{\mu_g}{\mu} \right) \quad (7)$$

The first term of equation (7) is a weighted sum of concentrations of subpopulations. This is called the within-group contribution. The second term represents the concentration between groups (subpopulations). This is the so-called between-group contribution. If the averages of all groups are all the same then this term vanishes.

We, finally, note that the number of sources was *not* kept fixed. The proposed measures all satisfy the so-called replication axiom. This means that the concentration or evenness of vectors such as (1,2,3), (1,1,2,2,3,3) and (1,1,1,2,2,2,3,3,3) are all the same. This is clear as such vectors all have the same Lorenz curve.

Values of these concentration measures for the data in Tables A,B,C and D are given in the Appendix.

### **A second approach to concentration: comparison with a standard vector**

Sometimes one is not interested in the concentration of items over sources. One is merely interested in how different the concentration is with respect to a standard. This standard can be an internal or an external standard. If data are given in tabular form such as Table A or D then the internal standards are found in the last row (vector C for countries) or last column (vector J for journals). An external standard can, e.g., be the distribution of the population of different countries, which is to be compared with the publication output.

In the first case one is interested to know if the concentration of articles over countries for a particular journal is the same as that for all journals together; or one is interested to see if the distribution of journals over countries is the same as the global (publication) use of these journals. An example of the use of an internal standard is given in the Viles-French article (1999) where the authors study locality of a topic  $t$  over a distributed system by comparing the contribution of documents on topic  $t$  with the total distribution of documents in the system. In the case of an external standard one wants to know if this standard 'explains' the observed inequality. If a (relative) vector of observations coincides with the standard, this means that all inequality can be explained by this standard. An example of this is the case that differences in publications between countries are explained by differences in population. In both cases this leads to the use of weighted Lorenz curves and measures respecting these curves.

Weighted Lorenz curves are constructed as follows (Theil, 1967; Patil and Taillie, 1982; Rousseau, 1992a). Let  $S = (s_1, s_2, \dots, s_N)$  denote the standard vector and let  $X = (x_1, x_2, \dots, x_N)$  denote the distribution vector that we want to compare with this standard. Note that now indices must correspond. If, e.g.,  $X$

denotes numbers of publications and  $S$  denotes population then  $x_i$  and  $s_i$  must refer to the same country  $C_i$ . We assume, moreover, that none of the components of  $S$  is zero. In order to construct the Lorenz curve for comparisons with a standard the components of both vectors are ordered in such a way that

$$\frac{x_1}{s_1} \geq \frac{x_2}{s_2} \geq \dots \geq \frac{x_N}{s_N} \quad (8)$$

Next we normalise the vectors  $X$  and  $S$ , leading to vectors  $A_X$  and  $W$ , where

$$a_i = \frac{x_i}{\sum_{j=1}^N x_j} \quad \text{and} \quad w_i = \frac{s_i}{\sum_{j=1}^N s_j} \quad (9)$$

Note that normalizing does not change the order. Finally, the weighted Lorenz curve is defined as the broken line connecting the origin  $(0,0)$  to the points with components

$$\left( \sum_{j=1}^i w_j, \sum_{j=1}^i a_j \right)_{i=1, \dots, N} \quad (10)$$

For a fixed standard these Lorenz curves again introduce a partial order in the set of  $N$ -vectors.

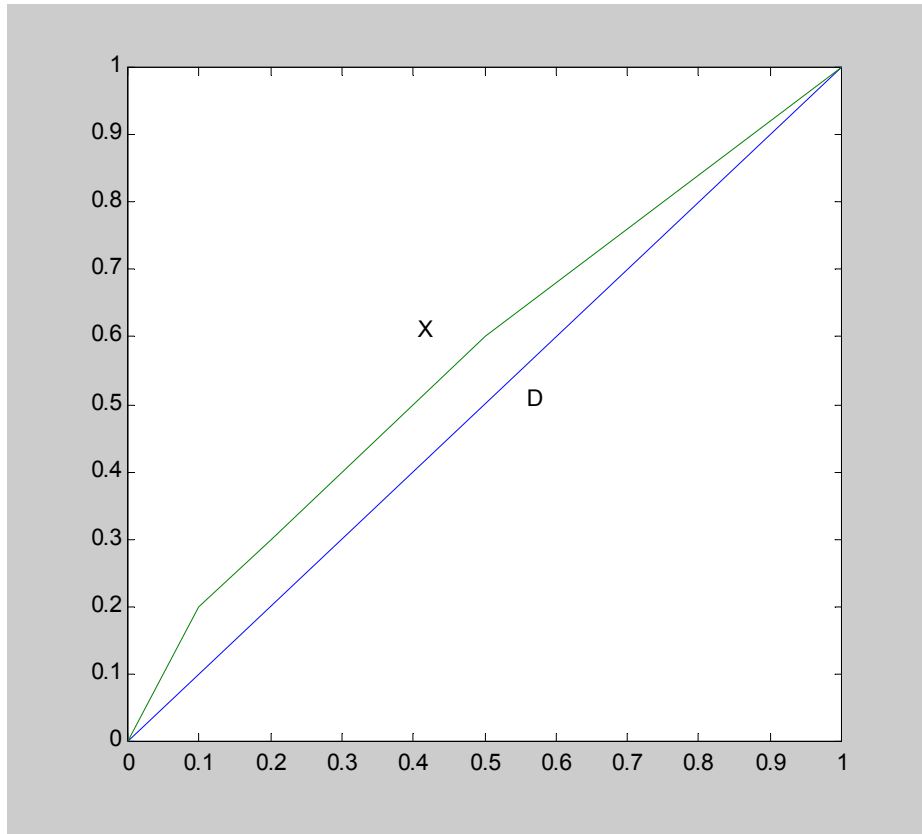


Fig. 2 Weighted Lorenz curve:  $X = (0.2, 0.1, 0.3, 0.4)$   
weighted with respect to  $(0.1, 0.1, 0.3, 0.5)$ .

Functions that respect this partial order are called measures of asymmetric relative concentration (Egghe & Rousseau, 2001a). The term 'relative' refers to the fact that one compares with a standard. The term 'asymmetric' stresses the fact that the roles of the standard and the vector under study cannot be interchanged. Examples of such measures are:

1) The asymmetric (or weighted) entropy measure:

$$H_w(X) = \sum_{i=1}^N a_i \ln \left( \frac{a_i}{w_i} \right) \quad (11)$$

2a) The asymmetric (or weighted) squared coefficient of variation:

$$V_w^2(X) = \sum_{i=1}^N \frac{(a_i - w_i)^2}{w_i} \quad (12)$$

2b) Another form of the weighted squared coefficient of variation, resembling more Simpson's measure (or the Herfindahl index):

$$V_w^2(X) = \left( \sum_{i=1}^N \frac{1}{w_i} a_i^2 \right) - 1 \quad (13)$$

3) The asymmetric (or weighted) Gini index:

$$G_w(X) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N |w_i a_j - w_j a_i| \quad (14)$$

The interpretation of this index is the same as that of the (unweighted) Gini index, namely twice the area between the Lorenz curve and the diagonal. This measure has been used in studies of the localization of industry (using internal, as well as external standards), under the name of 'locational Gini coefficients' (Krugman, 1991; Zitt et al., 1999). Krugman used an internal standard, while Zitt et al. used an external standard (population).

4) The length of the weighted Lorenz curve:

$$LOR_w(X) = \sum_{i=1}^N \sqrt{a_i^2 + w_i^2} \quad (15)$$

Proofs that these functions respect the partial order induced by weighted Lorenz curves are given in (Egghe, 2000; Egghe and Rousseau, 2001a).

Note that the studied vector and the 'standard' have in a natural way the same number of sources.

### **A third approach: source per source comparison of items: absolute differences.**

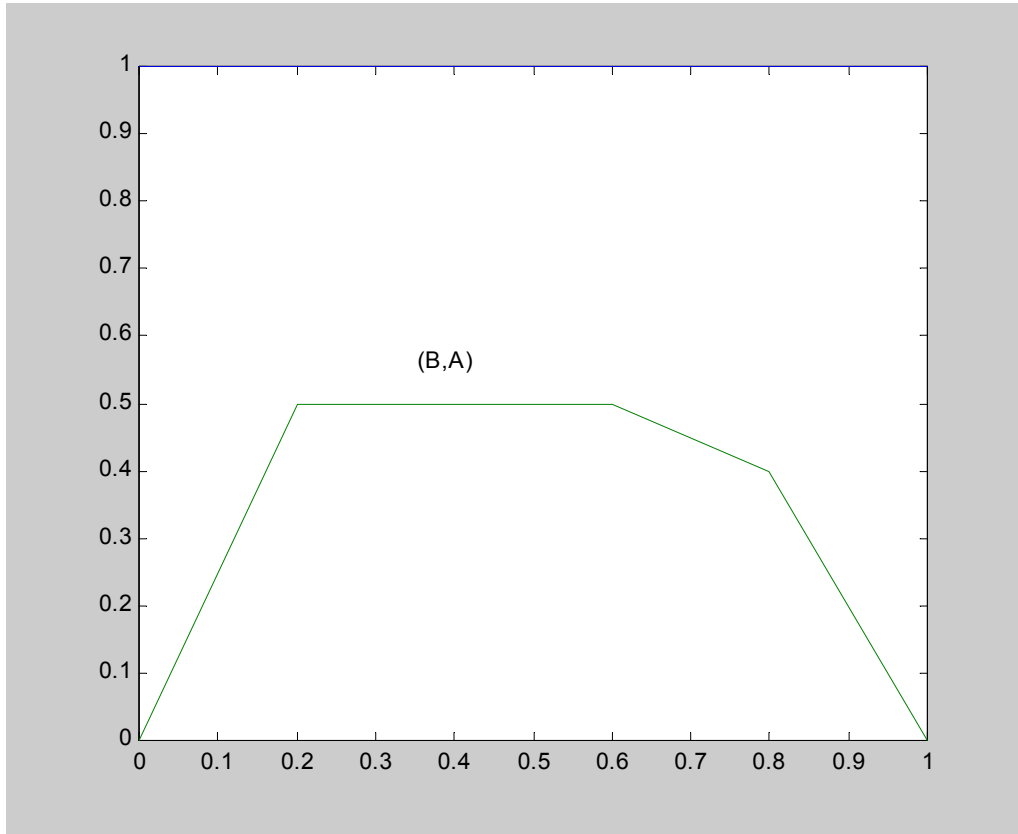
In this approach one directly compares relative vectors. Relative contributions of the same sources (but, e.g. at different times) are compared by taking differences. These differences can be positive or negative and one compares with the zero-vector. This means that absolute differences in relative contributions become important now.

Also here one can construct a kind of Lorenz curve. Because Professor Egghe was the first to propose this curve, we like to call it the Egghe-Lorenz curve. It is constructed as follows. Let  $X = (x_i)_{i=1, \dots, N}$  and  $Y = (y_i)_{i=1, \dots, N}$  be two  $N$ -vectors and let  $A = (a_i)_{i=1, \dots, N}$  and  $B = (b_i)_{i=1, \dots, N}$  denote their relative vectors (sum of all components equal to one). Then the components of the difference vector  $D = (d_i)_{i=1, \dots, N}$  with  $d_i = a_i - b_i$  are ranked from largest to smallest. Finally, putting

$$t_i = \sum_{j=1}^i d_j = \sum_{j=1}^i (a_j - b_j) \quad (16)$$

the Egghe-Lorenz curve is obtained by joining the origin (0,0) with the points with coordinates

$$\left( \frac{i}{N}, t_i \right)_{i=1, \dots, N} \quad (17)$$



**Fig. 3 The Egghe-Lorenz curve of (B,A), with  $B = (0.6, 0.2, 0, 0.1, 0.1)$  and  $A = (0.1, 0.2, 0, 0.2, 0.5)$**

Note that this curve always ends in the point (1,0). Similar to other Lorenz curves also this one leads to a partial order and functions that respect this partial order are the ones we are interested in. Such functions are known as Egghe's measures of symmetric relative concentration (Egghe, 1988, 1990; Egghe & Rousseau, 2001a). Here the term 'relative' again refers to the fact that one compares with a kind of standard. Indeed, one compares the difference vector with the all zero vector. Examples of such measures are:

$$C_r(X,Y) = -\frac{1}{N} \sum_{i=1}^N i d_i \quad (18)$$

where the  $d_i$  are ranked in decreasing order. This is nothing but the area under the Egghe-Lorenz curve. Another example is

$$V_r^2(X, Y) = N \sum_{i=1}^N d_i^2 \quad (19)$$

This is the adapted Simpson or Herfindahl index of the relative difference vector.

Similarly, one can use the length of the Egghe-Lorenz curve:

$$LOR_r(X, Y) = \sum_{i=1}^N \sqrt{d_i^2 + \frac{1}{N^2}} \quad (20)$$

It is also possible to use relative differences in a source per source comparison of items, but we will not discuss this here.

We end these sections on Lorenz curves and concentration measures by observing that the exact relation between different forms of Lorenz curves and the appropriate measures follows from a general mathematical theory studied by Egghe (2000).

## Application

As an application we have studied the inequality in publication output of the People's Republic of China's administrative units (provinces, autonomous regions and municipalities) over the period 1989-1998. Data are taken from the Chinese Science Citation Database and the Chinese Scientometric Indicators (Jin & Wang, 1999; Jin, 2000). For diverse reasons Hong Kong, Macao and Taiwan are not included. Chongqing became a municipality in 1997. In our data it is considered as a part of Sichuan, as it was before 1997 (hence, we have 30 regions in total). In this first example we have restricted ourselves to the coefficient of variation. We note that different acceptable inequality measures differ in the way they are sensitive to particular transfers. Table 1 shows the coefficient of variation for the publication data of China's administrative units.

**Table 1.** Year and inequality of publication output over the regions, as measured by the coefficient of variation.

| Year | Coefficient of variation |
|------|--------------------------|
| 1989 | 1.479                    |
| 1990 | 1.436                    |
| 1991 | 1.415                    |
| 1992 | 1.407                    |
| 1993 | 1.389                    |
| 1994 | 1.394                    |
| 1995 | 1.408                    |
| 1996 | 1.215                    |
| 1997 | 1.232                    |
| 1998 | 1.232                    |

It is clear that there is a large inequality in publication between China's administrative units. Figure 4 shows the change in publication inequality over the period 1989-1998. Generally, the inequality between the publication output of China's different regions decreases slowly. The update of CSCD's database in 1996 increased this process considerably. We have also calculated the coefficient of variation for the publication output of China's regions for the year 1998, according to the SCI (1.680) and according to CSTPC (1.003). The Gini index for this same year is 0.642 (SCI), 0.551 (CSCD) and 0.494 (CSTPC). These differences are quite remarkable.

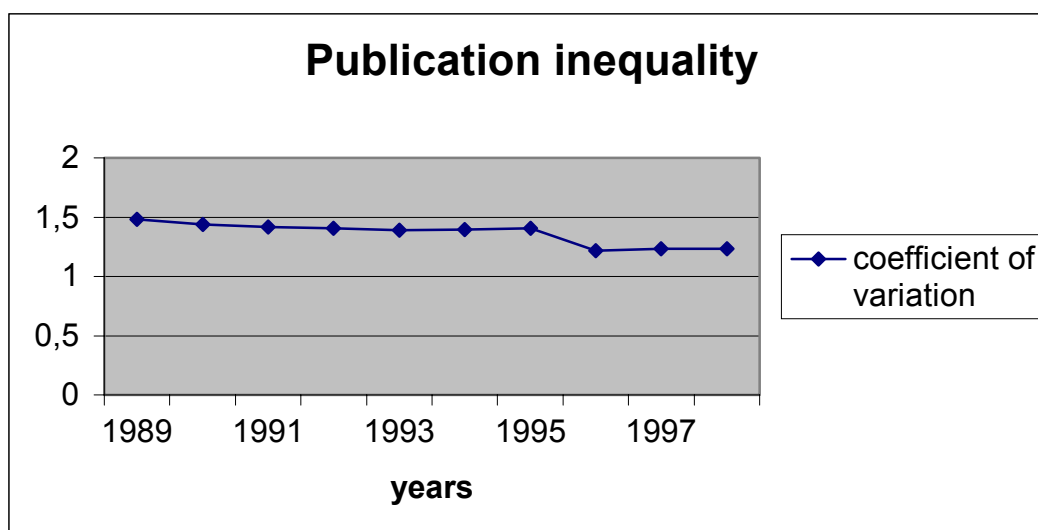


Fig. 4 Publication inequality as measured by the coefficient of variation (CSCD data)

The population inequality in this period as measured by the coefficient of variation is about 0.68. This number is considerably smaller than that of the publication inequality. Such a result, however, is not surprising. It has been

observed at several occasions (Allison, 1980; Rousseau, 1992b) that in general the inequality in 'use' is larger than the inequality in 'availability'. Examples of availability-use pairs are: publications and their citations; available CDs in a public library and the number of loans, and in our case: population and publications. It was suggested (Rousseau, 1992b) that the underlying mechanism for this phenomenon is a kind of positive reinforcement. Activities that are rewarded continue to be performed, while activities that are not rewarded tend to be stopped. Table 2 shows the inequality per administrative unit, over the years.

**Table 2.** Year and inequality of publication output over the years (per region), as measured by the coefficient of variation

| Regions           | Coefficient of variation | Rank |
|-------------------|--------------------------|------|
| Anhui             | 0.401                    | 24   |
| Beijing           | 0.320                    | 25   |
| Fujian            | 0.435                    | 17   |
| Gansu             | 0.228                    | 30   |
| Guangdong         | 0.608                    | 4    |
| Guangxi<br>Zhuang | 0.669                    | 3    |
| Guizhou           | 0.477                    | 13   |
| Hainan            | 0.411                    | 21   |
| Hebei             | 0.503                    | 11   |
| Heilongjiang      | 0.557                    | 6    |
| Henan             | 0.502                    | 12   |
| Hubei             | 0.449                    | 15   |
| Hunan             | 0.677                    | 2    |
| Inner<br>Mongolia | 0.768                    | 1    |
| Jiangsu           | 0.431                    | 18   |
| Jiangxi           | 0.530                    | 9    |
| Jilin             | 0.278                    | 29   |
| Liaoning          | 0.316                    | 26   |
| Ningxia Hui       | 0.560                    | 7    |
| Qinghai           | 0.316                    | 27   |
| Shaanxi           | 0.417                    | 20   |
| Shandong          | 0.541                    | 8    |
| Shanghai          | 0.304                    | 28   |
| Shanxi            | 0.601                    | 5    |
| Sichuan           | 0.404                    | 23   |
| Tianjin           | 0.420                    | 19   |
| Tibet =<br>Xizang | 0.458                    | 14   |
| Xinjiang<br>Uygur | 0.405                    | 22   |
| Yunnan            | 0.515                    | 10   |
| Zhejiang          | 0.443                    | 16   |

Inner Mongolia has the largest inequality of publication output over the years, while Gansu and Jilin are the most stable regions. Note that the publication output of the big cities Beijing and Shanghai is also very stable over the years. Not surprisingly we see that Guangdong's relative increase in publications puts it at rank 4.

As an example of the use of an external standard we have calculated  $V_w^2$  with respect to population (external standard) for the year 1998. This yields the value  $V_w^2 = 5.59$  for data based on the CSCD, 3.39 for data based on the CSTPC, and 11.19 for data based on the SCI.

Table 3 gives values of  $V_r^2$  a measure of symmetric relative concentration, comparing consecutive years (CSCD data).

**Table 3.** Values of  $V_r^2$ , comparing publications in consecutive years (CSCD data).

| Period | $V_r^2$ | Rank |
|--------|---------|------|
| 89-90  | 0.0068  | 2    |
| 90-91  | 0.0021  | 6    |
| 91-92  | 0.0052  | 3    |
| 92-93  | 0.0021  | 6    |
| 93-94  | 0.0033  | 5    |
| 94-95  | 0.0046  | 4    |
| 95-96  | 0.0685  | 1    |
| 96-97  | 0.0021  | 6    |
| 97-98  | 0.0015  | 9    |

Table 4 clearly shows the effect of CSCD's update in the year 1996. It seems further that relative changes have been reduced in recent times. The 95-96 value of this measure for symmetric relative concentration shows its sensitivity. We consider this observation as a pro for the use of this measure.

We note that evenness measures can be used to define the notion of a core in a mathematically precise way (Egghe & Rousseau, 2001b).

## Conclusion

It is important to be precise in stating the aim of concentration and evenness measurements. It might often be useful to measure different aspects of concentration. Different forms of concentration (and its opposite, evenness) are important indicators that have been neglected too much in the past (Rousseau, 1998). Such indicators are not only useful for people interested in

science management, but also for people interested in the sociological structure of the scientific community.

Counting publications is a first step in an evaluation exercise, counting citations a second one. Impact, or citations per publications is clearly a higher level of evaluation. Using concentration or diversity measures yields a macro view on the allocation of funds and the resulting output. Assuming that a number of research groups receive the same input, then the resulting inequality measures for their outputs are valid indicators of the total unbalance in the system. Assuming on the other hand, an unbalanced starting position, where groups or institutes have different initial conditions, then comparing the inequality between the input position and that of the output again yields a valid indicator about the performance of the system as a whole. We recall from our previous talk (Rousseau, 2000a) that the use of DEA (data envelopment analysis) is another approach to study input-output relations.

#### Acknowledgements.

I would like to thank Profs. Leo Egghe (LUC, Belgium) and Jin Bihui (DICCAS, Beijing) for help and encouragement during the preparation of this article. I further thank Dr. Wang Yan (ISTIC, Beijing) for sending me the CSTPC data for 1998.

#### References

- P.D. Allison (1978). Measures of inequality. *American Sociological Review*, 43, 865-880.
- C. Dagum (1980). The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée*, 33, 327-367.
- L. Egghe (1988). The relative concentration of a journal with respect to a subject and the use of online services in calculating it. *Journal of the American Society for Information Science* 39, 281-284.
- L. Egghe (1990). A new method for information retrieval based on the theory of relative concentration. *Proceedings of the 13<sup>th</sup> International Conference on Research and Development in Information Retrieval (SIGIR)* (Vidick, ed.), Brussels, 469-493.
- L. Egghe (2000). Construction of concentration measures for general Lorenz curves using Riemann-Stieltjes integrals. Preprint.
- L. Egghe and R. Rousseau (2001a). Symmetric and asymmetric theory of relative concentration and applications. *Scientometrics* (to appear).
- L. Egghe and R. Rousseau (2001b). The core of a scientific subject: an exact definition using fuzzy set theory (work in progress).
- B. Jin (2000). The development of *Chinese Scientometric Indicators*. Paper presented at the Second Berlin Workshop on Scientometrics and Informetrics/ Collaboration in Science and Technology, 1-4 September 2000.
- B. Jin and B. Wang (1999). Chinese Science Citation Database: its construction and application. *Scientometrics*, 45, 325-332.

- A.R. Khan and C. Riskin, C. (1998). Income and inequality in China: composition, distribution and growth of household income, 1988 to 1995. *China Quarterly*, 154, 221 - 253.
- P. Krugman (1991). *Geography and Trade*. Leuven: University Press.
- M.O. Lorenz (1905). Methods of measuring concentration of wealth. *Journal of the American Statistical Association* 9, 209-219.
- D. Nijssen, R. Rousseau and P. Van Hecke (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses* 13, 33-38.
- G.P. Patil and C. Taillie (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Society*, 77, 548-561.
- R. Rousseau (1992a). *Concentration and diversity in informetric research*. Ph.D. thesis, University of Antwerp (UIA).
- R. Rousseau (1992b). Concentration and diversity of availability and use in information systems: a positive reinforcement model. *Journal of the American Society for Information Science*, 43, 391-395.
- R. Rousseau (1998). Evenness as a descriptive parameter for department or faculty evaluation studies. In: *Informatiewetenschap 1998* (De Smet, ed.). Antwerpen: Werkgemeenschap Informatiewetenschap, 135-145.
- R. Rousseau (2000a). Bibliometric and econometric indicators for the evaluation of scientific institutes (in Chinese). In: *R&D Evaluation and Indicators*, (Jiang Guo-Hua, ed.), Beijing: Red Flag Publishing House, 16-37.
- R. Rousseau (2000b). Concentration measures as scientometric indicators. *Book of Abstracts. Sixth International Conference on Science and Technology Indicators* (Leiden, 24-27 May 2000), p. 92-93.
- R. Rousseau and P. Van Hecke (1999). Measuring biodiversity. *Acta Biotheoretica* 47, 1-5.
- Shorrocks, A.F. (1980). The class of additively decomposable inequality measures. *Econometrica*, 48, 613-625.
- Theil, H. (1967). *Economics and information theory*. Amsterdam: North-Holland.
- C.L. Viles and J.C. French (1999). Content locality in distributed digital libraries. *Information Processing and Management* 35, 317-336.
- M. Zitt, R. Barré, A. Sigogneau and F. Laville (1999). Territorial concentration and evolution of science and technology activities in the European Union: a descriptive analysis. *Research Policy*, 28, 545-562.

## Appendix

## Values of concentration measures for Table A

## Total

|             |       |
|-------------|-------|
| Gini        | 0.495 |
| Theil       | 0.439 |
| Mod.Simpson | 2.160 |
| Variation   | 1.079 |
| Length      | 1.539 |

## Per row (journal)

| measure   | J1    | J2    | J3    |
|-----------|-------|-------|-------|
| Gini      | 0.400 | 0.450 | 0.433 |
| Theil     | 0.292 | 0.413 | 0.349 |
| Mod S     | 1.540 | 1.992 | 1.804 |
| Variation | 0.735 | 0.995 | 0.897 |
| Length    | 1.513 | 1.542 | 1.522 |

## Per column (country)

| measure   | C1    | C2    | C3    | C4    |
|-----------|-------|-------|-------|-------|
| Gini      | 0.267 | 0.267 | 0.200 | 0.300 |
| Theil     | 0.155 | 0.148 | 0.069 | 0.162 |
| Mod S     | 1.260 | 1.320 | 1.140 | 1.305 |
| Variation | 0.510 | 0.566 | 0.374 | 0.552 |
| Length    | 1.470 | 1.464 | 1.438 | 1.477 |

Note the different ranking according to the Gini, length and the Theil on the one hand, and the Simpson and the variation coefficient on the other hand. This can be explained by the fact that the corresponding Lorenz curves cross.