THE BRADFORD DISTRIBUTION AND THE GINI INDEX

Q.L. BURRELL

Statistical Laboratory, Department of Mathematics, University of Manchester,
Oxford Road, Manchester, M13 9PL (UK)

(Received March 12, 1990)

It is pointed out that the so-called "Bradford distribution" derived by Leimkuhler is more properly viewed as the theoretical form of a variant of the Lorenz curve. The equation of this Leimkuhler curve allows an easy calculation of the Gini coefficient of concentration which can be compared with empirical values.

1. Introduction

In what was perhaps the founding paper on bibliometrics, *Bradford*[1,2] considered the problem of describing the manner in which references in a bibliography of a particular subject are "scattered" (distributed) over the journals in which they appear and presented two (non-equivalent) versions of a "law of scattering". In developing Bradford's method of analysis further, *Leimkuhler*[3] derived a mathematical form for what he termed the Bradford distribution depending upon a single parameter interpretable from the so-called verbal formulation of Bradford's law.

In this note we first point out that Leimkuhler's approach is essentially equivalent to the construction of a Lorenz curve for the bibliographic data. We then consider the particular functional form proposed for this curve and show that it allows an easy calculation of the Gini index or coefficient of concentration. Finally we look at some data sets in order to compare the theoretical value with the empirical value derived from the data.

2. The Leimkuhler curve

In analyzing the distribution of references over journals in a bibliography it is as if we have a population of N (say) productive journals for each of which we are able to observe the value of the random variable X where

X = no. of references produced by the journal.

Elsevier, Amsterdam - Oxford - New York - Tokyo
Akadémiai Kiadó

If we write

$$f(j) = \text{no. of journals producing } j \text{ references}, \; j = 1, 2, \ldots$$

then

$$P(X=j) = f(j)/N \quad , \quad j = 1, 2, \ldots .$$

Bradford's original approach[1,2] did not proceed with the usual statistical consideration of the distribution of X but instead focussed on the ranks of the journals, arranged in decreasing order of productivity. The ranks of the journals are thus defined as

$$\begin{aligned} r(j) &= \text{rank of a journal carrying } j \text{ references} \\ &= \text{no. of journals carrying at least } j \text{ references} \\ &= f(j) + f(j+1) + \ldots \\ &= \sum_{i \geq j} f(i) \quad , \quad j = 1, 2, \ldots\ldots \end{aligned}$$

or, in terms of what might be termed the *proportional rank*,

$$\begin{aligned} r(j)/N &= \sum_{i \geq j} f(i)/N \\ &= P(X \geq j) \\ &= \bar{F}(j) \, , \quad \text{say,} \quad j = 1, 2, \ldots . \end{aligned} \qquad (1)$$

This function F in statistical parlance would be termed the *tail distribution function* of X or, in some circumstances, the *survivor function* of X.

Note that F(j) just gives the proportion of journals carrying at least j references. Leimkuhler's analysis seeks to relate this to the proportion of the references in the bibliography carried by these journals. Thus let us write

$$R(j) = \text{total number of references carried by journals producing at least j references each}$$

$$= jf(j) + (j+1)f(j+1) + \ldots$$

$$= \sum_{i \geq j} if(i)\,, \quad \text{for} \quad j = 1, 2.... \quad .$$

If the total number of references in the bibliography is M then R(j)/M gives the proportion of total productivity accounted for by those journals producing j or more references. Adapting Leimkuhler's presentation, if $x=r(j)/N$ denotes "the fraction of documents in a collection which are most productive, $0 \leq x \leq 1$", then $G(x) = R(j)/M$ denotes "the proportion of total productivity contained in the fraction x" (quotations from Ref. 3). Thus *Leimkuhler* interpreted G as the cumulative distribution function of the most productive proportion of journals and, indeed, considered the corresponding probability density function and mean of the distribution. However, this is not strictly a legitimate statistical interpretation since the quantity whose distribution is being considered is not a random variable in the standard sense. Essentially, the problem is that the rank of a journal is not an observable i.e. it cannot be determined by consideration of the journal itself but only in conjunction with knowledge of the entire population.

If instead we work with the previously introduced random variable X, and if we put

$$\mu = M/N = \text{total no. of references/total no. of journals}$$

$$= \text{mean no. of references per journal}$$

$$= E[X],$$

then

$$R(j)/M = \sum_{i \geq j} if(i)/M$$

$$= \sum_{i \geq j} if(i)/N/(M/N)$$

$$= \sum_{i \geq j} iP(X=i)/\mu$$

$$= \Psi(j) \quad , \quad \text{say,} \quad j = 1, 2, ... \qquad (2)$$

We may thus describe Ψ as the (proportional) *tail moment function* of X.

The plot of points $(\bar{F},\Psi)$ may be considered for any random variable X having finite mean and we term it the *Leimkuhler plot* (Since X is discrete we only have the distinct points $(\bar{F}(j), \Psi(j))$, $j = 1, 2, ...,$ but it is often convenient to picture them as lying on a continuous curve, which we call the *Leimkuhler curve*).

If the journals had instead been ordered in increasing order of productivity we would similarly have been led to consider

$$F(j) = P(X \leq j)$$

and

$$\Phi(j) = \sum_{i \leq j} iP(X=i)/\mu, \quad j = 1, 2, ... \; .$$

The plot of Φ as ordinate against F as abscissa gives the so-called Lorenz curve of concentration (see e.g. *Kendall* et al.[4], p. 60) which is well-known in the field of economics, particularly in the context of income distributions. (In bibliometrics, *Burrell*[5], has pointed out that essentially the same sorts of curve are exploited by *Trueswell*[6-8] in his studies of monograph circulations in libraries.) The simple relationship between Leimkuhler and Lorenz curves is revealed by noting that

$$F(j) = 1 - P(X \geq j+1)$$

$$= 1 - \bar{F}(j+1)$$

and

$$\Phi(j) = 1 - \sum_{i \geq j+1} iP(x=i)/\mu$$

$$= 1 - \Psi(j+1)$$

with the graphical equivalence illustrated in Fig. 1. Note that it is a simple matter to show (cf Ref. 4, p.61) that the Leimkuhler curve is necessarily concave to the $\bar{F}$-axis.

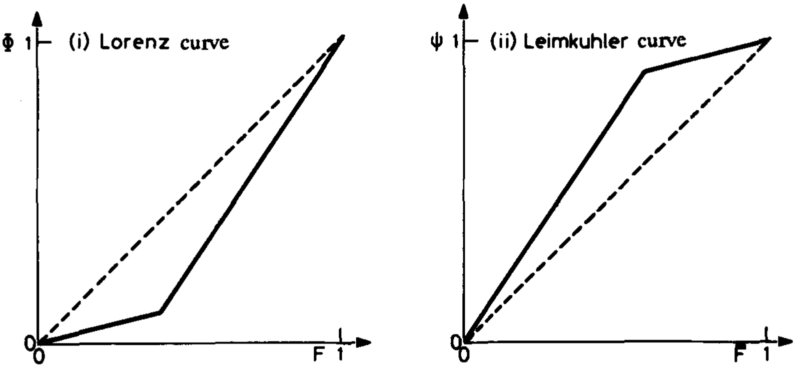

Fig. 1. Graphical illustration of the equivalence of the Lorenz and Leimkuhler curves

3. Theoretical considerations

3.1. The Bradford "distribution"

Starting from the so-called verbal formulation of Bradford's law of scattering[1,2], *Leimkuhler*[3] showed that the analytic form of G is

$$G(x) \quad = \ln(1+\beta x)/\ln(1+\beta) \ , \ 0 \leq x \leq 1, \qquad (3)$$

where $\beta > 0$ is a parameter which is characteristic of the particular bibliography. It is this function which *Leimkuhler* termed the *Bradford distribution*. (Actually, it has recently been pointed out by *Burrell*[9] that Leimkuhler's derivation implicitly assumes a stronger form of Bradford's law. This will not concern us here. We take the view that Eq. (3) provides a reasonable description of certain data sets and has some theoretical justification.) However, in view of the discussion in Section 2 this is something of a misnomer. In terms of our random variable X, Leimkuhler's derivation gives the functional relationship between $\bar{F}$, the tail distribution function of X and Ψ, the tail moment function of X, as

$$\Psi \ = \ln(1+\beta\bar{F})/\ln(1+\beta) \quad , \quad 0 \leq \bar{F} \leq 1. \tag{4}$$

Examples of this theoretical form of the Leimkuhler curve are given in Fig. 2 for various values of β.

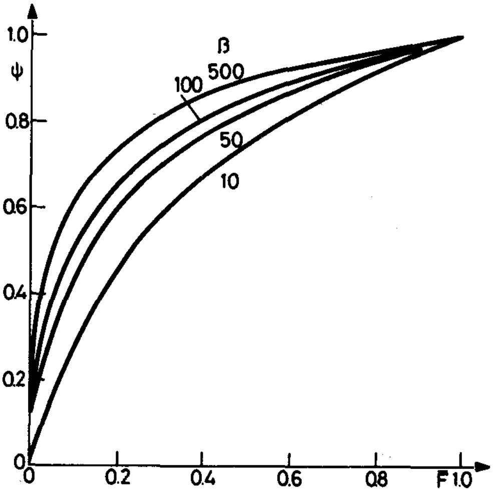

Fig. 2. The theoretical Leimkuhler curve $\Psi = \ln(1 + \beta\bar{F})/\ln(1 + \beta)$

If we are to consider whether a given discrete data set is well approximated by the Leimkuhler curve given by Eq. (4), and as we have noted this continuous curve really relates to a continuous probability distribution, we need to estimate the parameter β. In his original presentation, Leimkuhler viewed Eq. (3) (or Eq (4)) as a probability distribution and estimated β by the method of moments, which requires an iterative (Newton-Raphson) calculation. Although this does lead to a valid estimate, it is not particularly enlightening. Indeed *Brookes*[10] complained "nor is the parameter β simply related to any obvious characteristic of the journal collection". However, this last objection overlooks the derivation of the Leimkuhler curve (4) from the verbal form of Bradford's law.

A careful discussion of this derivation is given in Ref. 9, though for our purposes the following will suffice. Suppose that we split our ranked collection of journals into g equally-productive groups i.e. so that each group of journals produces the same number of references. Then Bradford's law states that the number of journals in the succeeding groups will be in the ratio $1 : b : b^2 : ... : b^{g-1}$, where the constant $b=b(g) > 1$ is called the Bradford multiplier. If this is the case (i.e. if Bradford's law holds for g groups) then the Leimkuhler plot of the (upper) proportional ranks of the g groups against the corresponding proportional productivities lie on the Leimkuhler curve (4) with $\beta = b^g\text{-}1$. If all choices of g lead to plots lying on the same Leimkuhler curve (and this is not necessarily the case[9]) then of course g is arbitrary, b^g is a characteristic for the collection (see *Egghe*[11]) and hence we may find an estimate for β by dividing the journals into groups à la Bradford and setting $\beta = b^g\text{-}1$. This simple relationship seems to satisfy Brookes' complaint.

3.2. The Gini index

In many fields, most notably in economics, measures have been proposed to measure the concentration (or dispersion) of a distribution. In order to compare concentrations in different populations it is important that such measures are independent of the particular scale of measurement adopted. (For a comprehensive discussion of such measures, and consideration of various desirable properties, within the field of bibliometrics, see *Egghe* and *Rousseau*[20]).

One measure which has been used extensively, e.g. in income distribution, is the Gini index, or coefficient of concentration. If Y is the random variable the concentration of whose distribution is being measured, then the Gini index is denoted γ_Y and is most succinctly expressed as

$$\gamma_Y = E[|Y_1\text{-}Y_2|]/2E[Y] \tag{9}$$

where Y_1, Y_2 are independent copies of Y. In terms of the distribution function F_Y we may write

$$\gamma_Y = \int\int | y_1\text{-}y_2| \, dF_Y(y_1)\, dF_Y(y_2)/2 \int y \, dF_Y\,(y)$$

Thus the Gini index is the mean (absolute) difference, divided by twice the mean and satisfies $0 \leq \gamma \leq 1$.

Kendall et al.[4] (p.61-62) show that the Gini index is equal to twice the area between the Lorenz curve and the line $\Phi = F$. Inspection of Fig. 1 then demonstrates:

Lemma. The Gini index is given by

$$\gamma = 2\,(\text{area beneath Leimkuhler curve}) - 1 \qquad (10)$$

From this it is a straightforward exercise in integration to show:

Corollary 1. For a probability distribution having the Leimkuhler curve given by (4), the Gini index is

$$\gamma = 1 - 2\,[(\ln(1+\beta))^{-1} - \beta^{-1}] \qquad (11)$$

A plot of the Gini index, γ, against the Leimkuhler parameter, β, is given in Fig 3 for values of β up to 10^5 (note that β is plotted on a logarithmic scale). It follows from (11) that as $\beta \to \infty$, $\gamma \to 1$ although, as can be seen from Fig. 3 the convergence is rather slow. In fact for very large values of β we can ignore the term β^{-1} in (11) and get the approximate formula

$$\beta \approx \exp 2/(1-\gamma)\,.$$

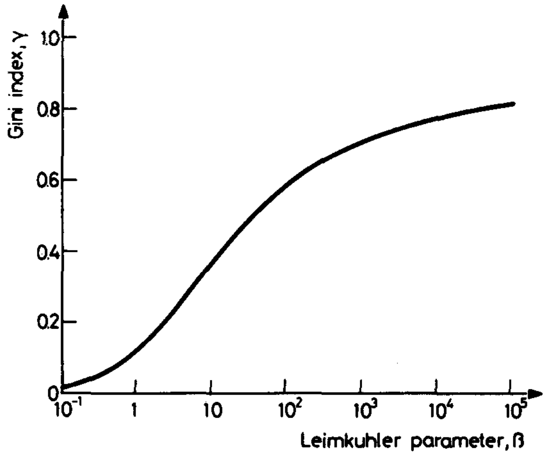

Fig. 3. The Leimkuhler parameter and the Gini index

Using this we find that to get γ as high as 0.9 for instance, requires $\beta \approx 4.85 \times 10^8$. We shall see in Section 4 that reported empirical studies suggest β to be most usually less than 500 with consequent values of γ less than about 0.68.

3.3 Calculation of the Gini index for empirical data

The following is no doubt well-known in the field of econometrics but, as it does not seem to have appeared in the bibliometric literature in this form, we include it here for the sake of completeness. Bearing in mind typical bibliometric contexts, and yet wishing to retain some generality, we suppose that the distribution of the random variable X is given by

$$P(X = j) = p_j, \quad \text{for } j = 1, 2, ..., m \ \ (< \infty).$$

Then, retaining our earlier notation, we have the following:

Theorem.

$$\gamma_X = 1 - \sum_{j=1}^{m} \bar{F}(j)^2/\mu_X \tag{12}$$

where $\mu_X = E[X]$.

Proof. Let X_1, X_2 be independent copies of X and consider $E[|X_1-X_2|]$.

Note first that $$P(|X_1-X_2| = j) = \begin{cases} P(X_1 = X_2)\cdot & , \text{if } j = 0, \\ 2P(X_1 - X_2 = j) & , \text{if } j = 1, 2, ..., m\text{-}1. \end{cases}$$

Thus $$P(|X_1-X_2| = 0) = \sum_{k=1}^{m} P(X_1 = X_2 = k) = \sum_{k=1}^{m} p_k^2,$$

while if $j \neq 0$, $$\{X_1-X_2 = j\} = \bigcup_{k=j+1}^{m} \{X_1=k\} \cap \{X_2=k-j\}$$

so that $$P(|X_1-X_2| = j) = 2\sum_{k=j+1}^{m} p_k\, p_{k-j}, \text{ for } j = 1, 2, ..., m\text{-}1$$

$$\begin{aligned} \therefore\ E[|X_1-X_2|] &= \sum_{j=1}^{m-1} jP(|X_1-X_2| = j) \\ &= 2\sum_{j=1}^{m-1} j \sum_{k=j+1}^{m} p_k\, p_{k-j} \\ &= 2\sum_{k=2}^{m} p_k \sum_{j=1}^{k-1} j\, p_{k-j} \\ &= 2\sum_{k=2}^{m} p_k \sum_{j=1}^{k-1} P(X \le j) \\ &= 2\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} p_k\, P(X \le j) \end{aligned}$$

$$= 2 \sum_{j=1}^{m-1} P(X \geq j+1)\, P(X \leq j)$$

$$= 2 \sum_{j=1}^{m-1} \bar{F}(j+1)\, [1-\bar{F}(j+1)]$$

$$= 2 \sum_{j=1}^{m} \bar{F}(j)\, [1-\bar{F}(j)], \text{ since } \bar{F}(1) = 1,$$

$$= 2\, [\mu_x - \sum_{j=1}^{m} \bar{F}(j)^2],$$

since

$$\sum_{j \geq 1} P(X \geq j) = E[X]$$

for any non-negative integer-valued random variable X.

Hence from Eq. (9),

$$\gamma_x = E[\,|X_1 - X_2|\,]/2\mu_x$$

we have $\gamma_x = 1 - \sum_{j=1}^{m} \bar{F}(j)^2/\mu_x$ as required.

In view of the favoured way of presenting many bibliometric data sets in ranked form, as described in Section 2, it is convenient to give an alternative form as follows:

Corollary 2. In the notation of Section 2,

$$\gamma_x = 1 - (MN)^{-1} \sum_{j=1}^{m} r(j)^2 \tag{13}$$

Proof. This is a straightforward substitution for

$$\bar{F}(j) = r(j)/N \text{ and } \mu_x = M/N\,.$$

Note that to apply Eq. (13) we need r(j) for each value of j from 1 to m (the maximal observed value), whether or not the corresponding p_j is positive. In the terminology of journal productivity, we need r(j) whether or not there are any journals producing j references. This causes no problem provided we recall our definition that r(j) gives the number of journals carrying *at least* j references and this is well-defined *for all* j = 1, 2, ...,m.

4. Empirical studies

In this section we consider some well-known data sets which have recently been re-analysed by *Egghe*[12] in another investigation of the connections between *Bradford* and *Leimkuhler* presentations. Leimkuhler plots (and Leimkuhler curves) for these data may be found in Ref. 2, but note that *Egghe* plots R against r rather than the standardized form R/M against r/N as described in Section 2. This, of course, is simply a re-scaling of the axes. Note also that Egghe assesses the goodness-of-fit of the data using the Kolmogorov-Smirnov statistic but, in the light of the discussion of the correct interpretation of the Bradford 'distribution' given in Section 2, this is inappropriate.

4.1. Applied geophysics

This is one of the bibliographies analyzed by *Bradford*[1,2]. According to *Egghe*[11] if we form just 3 groups of journals (i.e. if $g=3$) then we find the Bradford multiplier to be 5.49 and (from Section 3.1) our estimate of β is

$$\beta \quad = 5.49^3 - 1 = 164.47.$$

On the other hand with $g = 4$ we get $b = 2.78$ and

$$\beta \quad = 2.78^4 - 1 = 165.04.$$

Given that (i) Bradford's law is at best an approximate, empirical law, (ii) the discrete nature of the data involves a certain arbitrariness in the division of journals into groups, and (iii) the Bradford multiplier is non-constant between groups so the quoted values are averages, the similarity between these values is remarkable.

Turning to the Gini index, we have the theoretical value calculated from (11) with $\beta=165$ as $\gamma_1=0.6205$ while the observed value calculated from the data, using (12) or (13), is $\gamma_2=0.6176$.

4.2. Lubrication

This is Bradford's other data set and is one of the smallest regularly considered, with only 395 references from 164 journals. Again from Ref. 11, we find $g=3$, $b=3.4$

and hence $\beta=38.30$ or $g=7$, $b=1.69$ and $\beta=38.37$. Taking $\beta=38.3$ we find $\gamma_1=0.5074$ compared with the calculated value $\gamma_2=0.4762$.

4.3 ORSA

These are the data on operational research presented by *Kendall*[13] based on a bibliography published by the Operations Research Society of America. From Ref. 11, with $g=4$ and $b=4.56$ we get $\beta=431.37$. The theoretical and observed values of the Gini index are, with $\beta=431$.

$$\gamma_1 = 0.6751, \quad \gamma_2 = 0.6857 .$$

4.4 Schistosomiasis

A bibliography covering the literature on schistosomiasis from 1852 to 1962 which was analyzed and reported by *Goffman* and *Warren*[14] is one of the most extensive to appear, not only in the time period covered but in both the number of journals (1738) and the number of references (9914) featured. This is an interesting example since it is held not to obey Bradford's law properly (see *Egghe*[11,12]).

According to Egghe's analyses, taking $g=9$ groups gives an (average) Bradford multiplier $b=2.03$ and hence $\beta=584.42$, while $g=16$ and $b=1.49$ gives $\beta=590.17$. Again the similarity of the two values for β is remarkable, even if Egghe's choices of groupings seem somewhat arbitrary. Talking $\beta=590$ gives the theoretical value for the Gini index $\gamma_1=0.6900$, while the value calculated from the data is $\gamma_2=0.7177$.

4.5 Other bibliographies

In Ref. 12 are included also bibliographies on mast cells[14], information science (*Pope*[15]) and statistical methods (*Sachs*[16]). The Leimkuhler parameters and both the theoretical and empirical values of the Gini index of these and the previous bibliographies are summarized an Table 1.

Table 1
Leimkuhler parameter and Gini index for some well-known data sets, from *Egghe*[11,12]

Bibliography	Leimkuhler parameter, β	Gini index	
		Theoretical, γ_1	observed, γ_2
Applied geophysics[1,2]	165	0.6205	0.6176
Lubrication[1,2]	38.3	0.5074	0.4762
Operational research[13]	431	0.6751	0.6857
Schistosomiasis[14]	590	0.6900	0.7177
Mast cells[14]	114	0.5960	0.6278
Information science[15]	463	0.6786	0.7591
Statistical methods[16]	113	0.5954	0.6795

4.6. General comments

The last four examples in Table 1 exhibit substantial differences in the two values of the Gini index, the observed value far exceeding the theoretical value calculated from the Leimkuhler parameter, β. The basic reason for the discrepancy in each case is that the data set does not really conform with Bradford's law, and hence is not well approximated by the theoretical form of the Leimkuhler curve (see Ref. 12 for a more detailed commentary). In particular, these sets feature the so-called "Groos-droop" which means that the numbers of low-productivity journals are somewhat smaller than would be expected if Bradford's law applied. Since it is these numbers which dominate the calculation of the empirical Gini index this has the effect of increasing its value.

5. Concluding remarks

For bibliographies conforming to Bradford's law, the calculation of β and the corresponding Gini index γ_1 are sensible procedures, in particular since they are numerical values which allow comparisons between bibliographies in a simple quantitative fashion. In other cases there seems to be no reason - except perhaps to those who believe Bradford's law to be immutable - to fit such data to an inappropriate theoretical Leimkuhler curve. In all cases a Leimkuhler plot, in standardized form, gives a convenient graphical presentation allowing visual comparison of different bibliographies while the empirical Gini index gives a numerical value to allow comparison of degree of concentration.

The use of the Gini-index as a measure of concentration in bibliometrics has previously been advocated, in a somewhat different form, by *Pratt*[17] (see also *Carpenter*[18] and *Egghe*[19]). For a discussion of concentration measures in general see the recent survey of *Egghe* and *Rousseau*[20].

References

1. S.C. BRADFORD, Sources of information on specific subjects, *Engineering*, 137 (1934) 85-86.
2. S.C. BRADFORD, *Documentation*, London, Crosby Lockwood, 1948.
3. F.F. LEIMKUHLER, The Bradford distribution ,*Journal of Documentation*, 23 (1967) 197-207.
4. M.G. KENDALL, A. STUART, J.K. ORD, *The Advanced Theory of Statistics*, Vol. 1, *Distribution Theory* (5th ed.), London, Griffin, 1987
5. Q.L. BURRELL, The 80/20 rule: library lore or statistical law, *Journal of Documentation*, 41 (1985) 24-39.
6. R.W. TRUESWELL, Determining the optimal number of volumes for a library's core collection, *Libri*, 16 (1966) 49-60.
7. R.W. TRUESWELL, Some behavioural patterns of library users: the 80/20 rule, *Wilson Library Bulletin*, 43 (1969) 458-461.
8. R.W. TRUESWELL, Growing libraries: who needs them? A statistical basis for the no-growth collection, In D. GORE, (Ed.), *Farewell to Alexandria: Solutions to Space, Growth and Performance Problems of Libraries*, Westport, Connecticut, Greenwood Press, 1976, pp.72-104.
9. Q.L. BURRELL, The non-equivalence of the Bradford, Leimkuhler and Lotka laws. Research Report, Statistical Laboratory, University of Manchester, 1990.
10. B.C. BROOKES, The derivation and application of the Bradford-Zipf distribution, *Journal of Documentation*, 24 (1968) 247-265.
11. L. EGGHE, A note on different Bradford multipliers, *Journal of the American Society for Information Science*, 41 (1990) 204-209.
12. L. EGGHE, Applications of the theory of Bradford's law to the calculation of Leimkuhler's law and to the completion of bibliographies, *Journal of the American Society for Information Science*, (to appear).
13. M.G. KENDALL, The bibliography of operational research, *Operational Research Quarterly*, 11 (1960) 31-36.
14. W. GOFFMAN, K.S. WARREN, Dispersion of papers among journals based on a mathematical analysis of two diverse medical literatures, *Nature*, 221 (1969) 1205-1207.
15. A. POPE, Bradford's law and the periodical literature of information science, *Journal of the American Society for Information Science*, 36 (1975) 207-213.
16. L. SACHS, *A Guide to Statistical Methods and to the Pertinent Literature*, Literatur zur angewandte statistik, Berlin, Springer-Verlag, 1986.
17. A.D. PRATT, A measure of class concentration in bibliometrics, *Journal of the American Society for Information Science*, 28 (1977) 285-192.
18. M.P. CARPENTER, Similarity of Pratt's measure of class concentration to the Gini index, *Journal of the American Society for Information Science*, 30 (1979) 108-110.
19. L. EGGHE, Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule, *Journal of the American Society for Information Science*, 38 (1987) 288-297.
20. L. EGGHE, R. ROUSSEAU, Elements of concentration theory, In: L. EGGHE, R. ROUSSEAU, (Eds), *Informetrics 89/90*, Elsevier, Amsterdam, 1990.