The core of scientific subjects : an exact definition using concetration theory and fuzzy set theory.

The core of a scientific subject: an exact definition using concentration and fuzzy sets

LEO EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
and UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
E-mail: leo.egghe@luc.ac.be

and

RONALD ROUSSEAU
KHBO, Industrial Sciences and Technology,
Zeedijk 101, B-8400 Oostende, Belgium
& UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
& LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
E-mail:ronald.rousseau@kh.khbo.be

Abstract

Determining the core of a field's literature, i.e. its 'most important' sources, has been and still is an important problem in bibliometrics. In this article an exact definition of a core of a bibliography or a conglomerate is presented. The main ingredients for this definition are: fuzzy set theory, Lorenz curves and concentration measures. If one prefers a strict delineation, the fuzzy core can easily be defuzzified. The method we propose does not depend on the subjective notion of 'importance'. It is, moreover, completely reproducible. The method and the resulting core is also independent of the mathematical function (Lotka, Zipf, Bradford, etc.) that may be used to describe the relation between the set of sources and that of items.

Keywords

Core, concentration theory, fuzzy set theory

Introduction

Consider a set of sources. These sources may or may not have produced a number of items. As a generic name for this framework we use the term 'conglomerate' (Egghe & Rousseau, 2000). A conglomerate is nothing but a 'generalized bibliography', but goes much farther than the term 'bibliography' implies. Examples of conglomerates include: scientific disciplines where the sources are published documents and the items references in these documents, the Dutch fiction literature where sources are published fiction books and items the words used in these books, the Internet where sources are web pages and items are web domains (.com, .org, .uk, .cn, .be, etc.), the scientific collaboration network as seen from Australia, where sources are Australian scientists, and items are the countries of co-authors of these scientists. Many more examples of the 'conglomerate' idea can easily be given.

Since Bradford (1934) documentalists are interested in the determination of core journals of a scientific domain. McCain (1997) pointed out that the existence of highly productive and highly cited core journal literatures underlies the effectiveness of the ISI citation databases. Hence, determining cores has been and still is one of the problems informetricians try to solve. Peritz (1984) noted that the term 'core' itself (a 'colloquial term' in her classification of terms) is of unknown origin, though the concept itself goes back to Bradford (1934). Bradford himself called it the 'nucleus' and solved the problem of finding a nucleus (or nuclear zone) by dividing bibliographies into three parts. These parts were obtained by applying what we nowadays refer to as 'Bradford's law'. The first one of these, the so-called nucleus, was referred to as "the core journals". Yet, Egghe (1990) has shown that the number of groups in which one may subdivide the bibliography is completely arbitrary. Thus

there is nothing special about a division in three groups, nor about the first group, for a division into more groups would result in another (smaller) first group (Rousseau, 1993).

Another approach to the 'core' problem went as follows. Graphs (on semi-logarithmic scales) of bibliographies were drawn. These were described as consisting of a linear part, preceded by a non-linear part. The non-linear part was then referred to as the core (Brookes, 1969). As there is no scientific method to delineate the linear part from the non-linear one, and as, moreover, the 'linear part' was not linear at all (but part of a curve approaching a (linear) asymptote, it is clear that this method is not reproducible and, moreover, yields a highly subjective core.

Many practically inclined bibliometricians such as White and McCain (1990) identify the notion of 'core journals' with the notion of 'most important' journals, and hence complain that the investigations of informetricians have not led to an unequivocal method to determine the core journals of a field. As a reaction to this complaint, we would like to make two observations. First, it is impossible to give a precise meaning to such a vague term as 'most important journals'. Secondly, it is better, in our opinion, to distinguish between the terms 'core of a conglomerate' and 'most important sources'. Every conglomerate is composed in a different way and for various purposes. Journals (as sources) may be ranked according to the number of publications on some subject (Bradford type bibliographies), but also according to the number of citations these journals receive over a certain period of time (for scientometric purposes). In both cases, there will be a core, at least in the intuitive sense of the word, but these cores will be different. Important sources are important for different reasons, not only because of the number of items they produce. A journal, for instance, can be considered important because of its impact factor, but also because it is published at an old and respectable university, or because of the international prestige of its editor.

Admitting the value of previous attempts, especially in view of practical applications, we are convinced that we are still in need of a more objective way to determine a core. Admitting further the vagueness of this notion, we will use fuzzy set theory to define a core. This will lead to a scientific method to obtain a core. We will further suggest a method to defuzzify this core, leading to a crisp, i.e. well-defined set of core sources.

The main theory that will lead us to the notion of a core is the theory of inequality (concentration, diversity, evenness). In this way, we link this article to the one we presented twelve years ago at the Second International Conference on Scientometrics and Informetrics, held in London (Ontario) (Egghe-Rousseau, 1990).

A short review of the 'core' literature

In this section we give a short, and necessarily, incomplete overview of cores that have been determined with different methods and for different purposes.

Lists of core journals have been drawn and studied for behavioural medicine (Slater & Slater, 1994), immunology (Arora & Pawan, 1995), renewable energy (Shukla, 1996), and many more other subjects. Terrence Brooks determined the core journals of a rapidly changing research front (superconductivity) using a technique derived from the Bradford curve (Brooks, 1989). McCain (1995) used a database filtering approach to determine core journals in biotechnology (a multidisciplinary, R&D-related field). She used ISI's databases and Pergamon's *Biotechnology Abstracts* as filters to extract the core and highly productive non-core journals. These journals were clustered and mapped based on their co-citation and subject heading profiles. The Chinese databases CSCD and CSI are used, among other things, to determine core institutions in the country (Jin et al., 2001). Egghe (1999) made a theoretical model of the influence of a core collection of journals on the ultimate development of a citation database.

Joswick and Stierman (1997) compared core lists of most frequently used journals by faculty and students of Western Illinois University. Their most interesting, although not surprising, finding was that these lists were very dissimilar, showing that, even within one academic library, consultation habits differ markedly between user groups.

Bonitz, Bruckner and Scharnhorst (1999) introduced a totally different kind of core: the Matthew core. The so-called 'Matthew core journals' are the journals where the bulk of the re-distribution effects of citations takes place. In other words, the Matthew core consists of those journals (Nature Science, Physical Review B, ...) where the largest differences can be observed between the expected and the observed number of citations of a country.

Core journals in a subject field are sometimes simply defined as those journals that belong to SCI's or SSCI's subject listing of that field. Yet, Rice et al. (1989) pointed out four problems with this approach. First, some potential members may be ignored (e.g. because SCI decided to list them in a different subject category). Second, the JCR 'core list' may include journals that practitioners in the field do not accept as important journals in that discipline. Third, journals may be listed in several 'cores' and finally, because the JCR does not define its criteria for core membership, the validity of this 'core' notion is uncertain.

Defining 'the' core of a bibliography or a conglomerate

Assume we have a conglomerate with N sources. Assume further that these sources are ranked in decreasing order of production and that x_i denotes the number of items produced by the i^{th} source, S_i, i = 1, …, N. This conglomerate is hence characterized by the N-vector $X = (x_1, x_2, \ldots, x_N)$. We denote by $X_i = (x_1, x_2, \ldots, x_i, 0, ..,0)$ the i^{th} partial N-vector. Although the vector X_i can mathematically be identified with the vector $(x_1, x_2, \ldots, x_i)$ their Lorenz curves are clearly different. For our purposes these vectors represent different conglomerates. The vector X_i may, e.g. represent a university department with N researchers, of which N-i have no publications, while $(x_1, x_2, \ldots, x_i)$ represents the publications of a department with i researchers. Given a conglomerate $X = (x_1, x_2, \ldots, x_N)$, we denote by L_i the (concave) Lorenz curve of X_i. For more details on the construction of a Lorenz curve, we refer the reader to (Egghe & Rousseau, 1990; Egghe, 2001).

It is clear that, for i = 1,… N-1, L_{i+1} is at no point situated strictly above L_i. Indeed, let $X_i = (x_1, x_2, \ldots, x_i, 0, ..,0)$ be the i^{th} partial N-vector (with (N-i) zeros), and let $X_{i+1} = (x_1, x_2, \ldots, x_i, x_{i+1}, 0, ..,0)$ be the $(i+1)^{th}$ partial N-vector (with (N-i-1) zeros), then the Lorenz curves L_i and L_{i+1} are constructed as follows. L_i connects the points:

$$(0,0),\left(\frac{1}{N},\frac{x_1}{\sum_{n=1}^{i}x_n}\right),\ldots,\left(\frac{i-1}{N},\frac{\sum_{n=1}^{i-1}x_n}{\sum_{n=1}^{i}x_n}\right),\left(\frac{i}{N},1\right),\cdots(1,1)$$

while L_{i+1} connects the points

$$(0,0),\left(\frac{1}{N},\frac{x_1}{\sum_{n=1}^{i+1}x_n}\right),\ldots,\left(\frac{i-1}{N},\frac{\sum_{n=1}^{i-1}x_n}{\sum_{n=1}^{i+1}x_n}\right),\left(\frac{i}{N},\frac{\sum_{n=1}^{i}x_n}{\sum_{n=1}^{i+1}x_n}\right),\left(\frac{i+1}{N},1\right),\cdots,(1,1)$$

It is now clear that for every $i \in \{1,\ldots,N-1\}$, , L_{i+1} is at no point situated strictly above L_i. Consequently, a good concentration measure C always leads to (Egghe, 2001):

$$C(X_{i+1}) < C(X_i)$$

C(X), the concentration of the complete conglomerate, is equal to $C(X_N)$ (and is, of course, smaller than any of the $C(X_i)$). We recall (see e.g. Nijssen et al., 1998; Egghe & Rousseau, 2001) that acceptable concentration measures for a given N-vector $Y = (y_1, y_2, \ldots, y_N)$ are:

- The Gini index (G)

This measure is twice the area between the Lorenz curve and the diagonal. It is defined as:

$$G(Y) = \frac{N+1-2\sum_{i=1}^{N} i a_i}{N}, \quad where \quad a_i = \frac{y_i}{\sum_{j=1}^{N} y_j}$$

- The entropy or Theil concentration measure

$$H(Y)=\ln(N)+\sum_{i=1}^{N} a_i \ln(a_i)$$

- The coefficient of variation

$$V(Y)=\frac{\sigma_Y}{\mu_Y}$$

where μ_Y is the mean and σ_Y the standard deviation of the vector Y.

- The modified Simpson concentration measure

$$S(Y)=N\sum_{i=1}^{N} a_i^2 - 1$$

One can show that this measure is actually nothing but V^2, the squared coefficient of variation. We prefer, however the following form.

- The normalized coefficient of variation:

$$NV(Y)=\frac{2}{\pi}\arctan(V(Y))$$

NV(Y) always yields a value between 0 and 1, hence the term 'normalized'.

Now, we define the core membership value of the i-th source, denoted as $m(S_i)$, (with respect to a fixed concentration measure C) as:

$$m(S_i)=\frac{C(X_i)-C(X)}{C(X_1)-C(X)}$$

As the values $C(X_i)$ decrease when i increases from 1 to N the core membership value decreases from 1 to 0. We note the fact that the user still has the freedom to choose a concentration measure. Economists and sociologists face a similar choice when studying income inequalities between countries or regions.

Examples

Figure 1, derived from Table 1, shows the fuzzy 'core set' for Bradford's Applied Geophysics, calculated using the Gini concentration measure. The most productive source belongs to the core with membership value 1.0; the source at the rank 69 belongs to the core with a membership value of 0.70; the source at rank 200 belongs to the core with a membership value of 0.34. The last source, the 326th, has membership value 0.00. Note that, strictly speaking, once there are many sources with the same production this fuzzy membership function should become a step function.

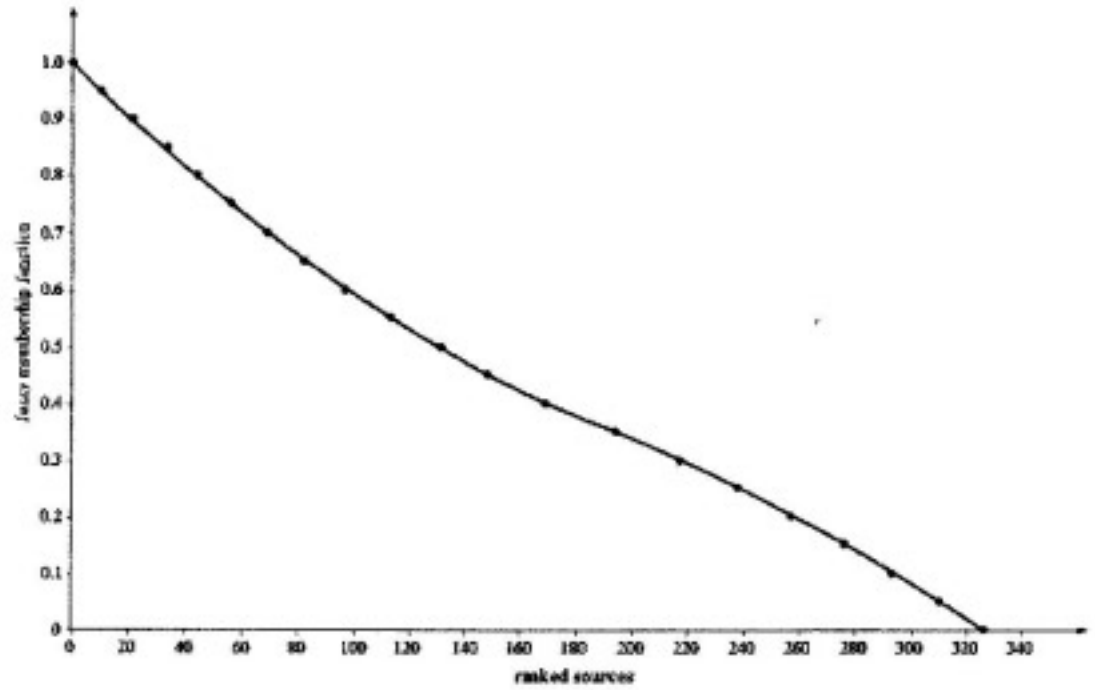

Figure 1. The fuzzy 'core set' for Bradford's Applied Geophysics, calculated using the Gini index

Table 1 Bradford's Applied Geophysics fuzzy membership values, calculated using the Gini index.

fuzzy membership value	rank of the source
1.00	1
0.95	10
0.90	21
0.85	33
0.80	44
0.75	56
0.70	69
0.65	82
0.60	97
0.55	113
0.50	131
0.45	148
0.40	169
0.35	194
0.30	217
0.25	238
0.20	257
0.15	276
0.10	293
0.05	310
0.00	326

Table 2 gives concentration values of a number of well-known bibliographic data sets. Most of these are studied in (Rousseau, 1994) to which we refer for further details. Exceptions are the conglomerates denoted as 'Sitations', 'Anthropology' and 'Penrose'. 'Penrose' refers to the author list of Penrose's famous book 'The emperor's new mind' (Penrose, 1990); 'Anthropology' refers to the 30-year cumulative author list for the journal Anthropological Linguistics (Jacobi & Propst, 1989); and 'Sitations' refers to the domain names with sites about 'informetrics' (Rousseau, 1997). Sets are ranked according to the value of the Gini index. The largest conglomerate contains 1011 sources (Pope's bibliography), the smallest 21 (small computer program).

Table 2 Real data: concentration values

Data set	Gini index	entropy or Theil measure	NV
Pope	0.7591	1.466	0.805
Rice a	0.7055	1.401	0.822
ORSA	0.6835	1.312	0.822
Sachs	0.6795	0.954	0.684
Sitations	0.6613	0.881	0.685
Mast cells	0.6278	0.862	0.694
Applied geophysics	0.6176	0.889	0.730
Rice b	0.6117	0.829	0.694
Nieuwenhuysen	0.5881	0.734	0.649
Rao, econ.	0.5865	0.656	0.627
Dresden	0.5858	0.710	0.664
Book index (Brookes)	0.5684	0.639	0.618
Lubrication	0.4762	0.487	0.582
Canadian authors	0.4560	0.442	0.570
Comp. Musicology	0.4483	0.486	0.620
Small computer program	0.4035	0.297	0.464
Informetrics	0.3411	0.382	0.591
Anthropology	0.2719	0.265	0.563
Schorr	0.2680	0.189	0.418
Murphy	0.2361	0.137	0.343
Penrose	0.2166	0.147	0.393
Radhakrishnan	0.1847	0.114	0.337
Kinnucan-Stat	0.1324	0.078	0.282

A suggestion for a well-determined (crisp) core

We define a p% core as the set of j most productive sources such that $m(S_j) \geq p$ with $m(S_{j+1}) < p$. Such a core depends on the used concentration measure. Ninety and ninety-five percent cores for the three concentration measures discussed earlier are shown in Table 3. For the entropy and the normalized coefficient of variation we also determined a 50% core. A comparison of these results will allow us to make a choice between these measures, and their corresponding cores.

Table 3. Number of sources forming a 95%, a 90 % core and a 50% core (the last one not determined for the Gini index): first number: 95% core, second one: 90% core, third one 50% core.

Data set	Gini index	Theil or entropy measure	NV
Pope	16-37	1-1-16	2-3-40
Rice a	4-7	1-1-8	1-1-11
ORSA	12-26	1-1-15	1-2-29
Sachs	3-6	1-1-7	1-2-12
Sitations	2-3	1-1-4	1-1-6
Mast cells	13-29	1-1-16	2-4-53
Applied geophysics	10-21	1-1-14	1-2-36
Rice b	2-6	1-1-7	1-1-12

Nieuwenhuysen	4-8	1-1-8	1-2-17
Rao, econ.	19-39	1-1-20	3-6-89
Dresden	8-17	1-1-13	2-3-36
Book index (Brookes)	3-6	1-1-8	1-2-17
Lubrication	6-11	1-1-10	1-3-29
Canadian authors	10-19	1-1-14	2-4-49
Comp. Music.	20-41	1-1-20	2-5-88
Small computer program	1-2	1-1-4	1-1-6
Informetrics	27-59	1-1-22	2-6-135
Anthropology	41-81	1-1-31	4-12-216
Schorr	15-31	1-1-17	3-6-98
Murphy	7-15	1-1-12	2-5-58
Penrose	11-23	1-1-15	3-6-78
Radhakrishnan	15-32	1-1-17	3-7-114
Kinnucan-Stat	11-21	1-1-14	3-6-85

The next table (Table 4) gives the average percentages and average number of sources included in the core (for the eight cases).

Table 4 Average percentage of the total number of sources included in the core

Gini 95% core	3.69%	(11.3 sources)
Gini 90% core	7.51%	(23.4 sources)
Entropy 95% = 90% core	0.73%	(1 source)
Entropy 50% core	6.01%	(13.6 sources)
NV 95% core	0.99%	(1.87 sources)
NV 90% core	1.65%	(3.91 sources)
NV 50% core	19.5%	(57.1 sources)

We propose using a 90% Gini core in practice. Indeed, it makes no sense to have a core that consists of only a few sources (the entropy and the normalized coefficient of variation 90 and 95% cores). Using a 50% core and the entropy or NV gives more acceptable numbers, but using a 50% core is counterintuitive to the notion of 'most important sources'. As the Gini index has a clear geometrical interpretation and seems to yield acceptable results we prefer the 90% Gini core. This core contains, on average 7.5% of all sources. For the examples studied here the (relatively) smallest 90% Gini core is *Pope*'s with 3.7 % of the sources; the (relatively) largest one is the author list of the journal *Anthropological Linguistics* (denoted as:"*Anthropology*"), containing 11.3% of all sources in the 90% Gini core. Yet, the reader is free to have a different preference, as the notion of 'a core' is in fact a fuzzy notion. Moreover, different applications may need different cores. Our approach offers this flexibility.

We note that the method proposed in this article is more satisfactory than the one based on truncated variable length vectors derived from X, such as $(x_1, x_2, \ldots, x_i)$ (Rousseau, 1992, 1993). Indeed, although in practice the sequence of such vectors often have increasing values for concentration measures, it is possible to find counterexamples to this statement (Rousseau, 1992). An approach based on a perfect Leimkuhler curve (Rousseau, 1987) is, from a theoretical point, even less satisfactory. This does not mean, of course, that it may yield acceptable results in practice.

Conclusion

In this article an exact definition of a core of a bibliography or a conglomerate has been presented. The main ingredients for this definition are: fuzzy set theory, Lorenz curves and concentration measures. If one prefers a strict delineation, the fuzzy core can easily be defuzzified. Our method has the advantage that it is independent of the fact whether or not the source-item relation in the conglomerate (generalized bibliography) can be described by a Leimkuhler or any other informetric distribution. Moreover, it is completely reproducible. Subjectiveness only enters through the used concentration measure and the, possible, choice of p (for the p% core).

References

J. Arora and U. Pawan (1995). Core journals in immunology: correlation analysis: rank v/s rank and rank v/s impact factor. *JISSI: The International Journal of Scientometrics and Informetrics*, 1(2), 83-97.

M. Bonitz, E. Bruckner and A. Scharnhorst (1999). The Matthew Index – concentration patterns and Matthew core journals. *Scientometrics*, 44, 361-378.

S. C. Bradford (1934). Sources of information on specific subjects. *Engineering*, 137, 85-86.

B.C. Brookes (1969). Bradford's law and the bibliography of science. *Nature*, 224, 953-956.

T. Brooks (1989). Core journals of the rapidly changing research front of "superconductivity". *Communication Research*, 16, 682-694.

L. Egghe (1990). 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*, 41, 469-492.

L. Egghe (1998). The evolution of core collections can be described via Banach space valued stochastic processes. *Mathematical and Computer Modelling*, 28, 11-17.

L. Egghe (2001). Construction of concentration measures for general Lorenz curves using Riemann-Stieltjes integrals. *Mathematical and Computer Modelling* (to appear).

L. Egghe and R. Rousseau (1990). Elements of concentration. In: *Informetrics 89/90* (L. Egghe & R. Rousseau, eds.) Elsevier, Amsterdam, 97-137.

L. Egghe and R. Rousseau (2000). Aging, obsolescence, impact, growth and utilization: definitions and relations. *Journal of the American Society for Information Science*, 51(11), 1004-1017.

L. Egghe and R. Rousseau (2001). Symmetric and asymmetric theory of relative concentration. *Scientometrics* (to appear).

K.P. Jacobi and K.B. Propst (1989). Anthropological Linguistics: thirty-year index (1959-1988). *Anthropological Linguistics*, 31, 3-52.

B. Jin, J. Zhang, D. Chen and X. Zhu (2001). Development of the *Chinese Scientometric Indicators (SCI)*. These *Proceedings*.

K.E. Joswick and J.K. Stierman (1997). The core list mirage: a comparison of the journals frequently consulted by faculty and students. *College & Research Libraries*, 58, 48-55.

K. McCain (1995). Biotechnology in context: a database-filtering approach to identifying core and productive non-core journals supporting multidisciplinary R&D. *Journal of the American Society for Information Science*, 46, 306-317.

K. McCain (1997). Bibliometric tools for serials collection management in academic libraries. *Advances in Serials Management*, 6, 105-146.

D. Nijssen, R. Rousseau and P. Van Hecke (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses*, 13, 33-38.

R. Penrose (1990). *The emperor's new mind.* Vintage: New York.

B.C. Peritz (1984). On the careers of terminologies; the case of bibliometrics. *Libri*, 34, 233-242.

R. E. Rice, C. L. Borgman, D. Bednarski and P.J.Hart (1989). Journal-to-journal citation data: issues of validity and reliability. *Scientometrics*, 15, 257-282.

R. Rousseau (1987). The nuclear zone of a Leimkuhler curve. *Journal of Documentation*.43, 322-333.

R. Rousseau (1992). *Concentration and diversity measures in informetric research*. Doctoral dissertation, University of Antwerp.

R. Rousseau (1993). Determination of a core of a bibliography. *Iaslic Bulletin*, 38, 49-57; 166a-166c.

R. Rousseau (1994). Bradford curves. *Information Processing and Management*, 30, 267-277.

R. Rousseau (1997). Sitations: an exploratory study. *Cybermetrics*, 1(1). http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

R. Rousseau (2000). Concentration and evenness measures as macro-level scientometric indicators. Paper presented at the Second International Symposium on Quantitative Evaluation of Research Performances and Sixth National Conference on Scientometrics and Informetrics (October 23-25, 2000, Shanghai, China).

M.C. Shukla (1996). Publication patterns in the field of renewable energy. An analysis of Indian Energy Abstracts. In: *Handbook of Libraries, Archives & Information Centres in India.* Volume 13 (B.M. Gupta, ed.). Segment Books: New Delhi, 309-328.

B. M. Slater and M. A. Slater (1994). Determining core journals in behavioral medicine. *Bulletin of the Medical Library Association*, 82, 70-72.

H.D. White and K. McCain (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24, 119-186.