Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 63, No. 2 (2005) 209–229

Bibliometric laws: Empirical flaws of fit

R. BAILÓN-MORENO,[a] E. JURADO-ALAMEDA,[a] R. RUIZ-BAÑOS,[b] J. P. COURTIAL[c]

[a] *Departamento de Ingeniería Química. Facultad de Ciencias, Universidad de Granada, Granada (Spain)*
[b] *Departamento de Biblioteconomía y Documentación, Facultad de Biblioteconomía y Documentación, Universidad de Granada, Granada (Spain)*
[c] *Laboratoire de Psychologie – Education – Cognition Développement (LabECD), Université de Nantes, Nantes (France)*

The bibliometric laws of Zipf, Bradford, and Lotka, in their various mathematical expressions, frequently present difficulties in the fitting of empirical values. The empirical flaws of fit take place in the frequency of the words, in the productivity of the authors and the journals, as well as in econometric and demographic aspects. This indicates that the underlying fractal model should be revised, since, as shown, the inverse power equations (of the Zipf–Mandelbrot type) are not adequate, as they need to include exponential terms. These modifications not only affect Bibliometrics and Scientometrics, but also, for the generality of the fractal model, apply to Economy, Demography, and even Natural Sciences in general.

Introduction

The present paper belongs to a series of papers related to the development of a Unified Scientometric Model. Here, the bibliometric laws and their empirical flaws of fit are analyzed.

The bibliometric laws of Zipf, Bradford, and Lotka are the pillars of Bibliometrics, Scientometrics and Informetrics. Given that Pareto's Law and Rule 80/20 are nothing more than extensions of the Lotka's Law to the fields of Economy and Demography, the Production Processes of Information, as generalized by Egghe and Rousseau, form part of the fundamentals of Social Sciences.

Another vital aspect to consider is the relationship of these laws with Fractal Mathematics, Chaos Theory, and Complex Systems. From the so-called lexicographic trees, Mandelbrot was able to find Zipf's Law of the frequency of words in language and, by similar arguments, its generalization to all types of social and natural phenomena: fluctuations in the stock market, population and wealth distribution, geometry of coastlines, plant structure, Brownian movement, surface structure of solids,

Received June 28, 2004

Address for correspondence:
RAFAEL BAILÓN-MORENO
Departamento de Ingeniería Química. Facultad de Ciencias, Campus de Fuentenueva
Universidad de Granada, 18701-Granada, Spain
E-mail: bailonm@ugr.es

0138–9130/US $ 20.00

atmospheric phenomena, etc.. All this implies that nature and society share a common substrate for which the mathematical expression is Zipf's Law together with its equivalents in Lotka and Bradford.

Before we continue, given that these will be the nucleus of the discussion in this article, it would be useful to review the mathematical expressions proposed for the bibliometric laws of Zipf, Bradford, and Lotka as well as their mathematical development in order to fit a greater number of empirical cases.

Zipf's Law

Let us consider a text of natural language and arrange the list of all the words that make up the language in descending order of frequency. The rank of a word denotes the position of this word in the aforementioned list. The simplest relationship that links the frequency of appearance and the rank is:

$$F = \frac{k_z}{R} \tag{1}$$

where F is the frequency of appearance of a word in a text, R is rank, and k_z is the Zipf constant. That is, the frequency is inversely proportional to the rank of the word. If we situate frequency and rank on the same member, the above expression takes on the following form:

$$FR = k_z \tag{2}$$

This could be stated as: the product of the frequency by the rank is a constant.

This law was proposed for the first time by the physicist E. U. Condon,[1] although it is currently known as Zipf's Law, after the linguist by that name published his famous book entitled *Human Behaviour and the Principle of Least Effort.*[2] This author arranged all the words of *Ulysses* by James Joyce in descending order of frequency and found the above relationship previously discovered by Condon. Thus, Eqs 1 and 2 should be called the Condon–Zipf Law.*

If we represent frequency against rank in a double-logarithmic diagram, we should get a straight line for which the ordinate at the origin is a logarithm of k_z with a slope equal to –1.

A noteworthy conclusion to be drawn from this law is that humans tend to prefer more usual words over rarely used ones. We are guided by the principle of least effort, which favours the common and discourages the uncommon.[2] In general, the most frequent words are also the shortest and easiest to pronounce.

* As one of our referees kindly called our attentiom, the first occurence of the idea in the literature can be found in *Gammes Sténographiques* (4th ed., Institute Stenographiques, Paris, 1916) by J. B. Estoup, therefore, the denomination Estoup–Zipf law is also justified.

Shortly after the studies of Zipf, it was confirmed that although frequency is always inversely related to rank, the distribution of words in a text usually veers, to a greater or lesser degree, from the standard or normal behaviour represented by Eqs 1 and 2. Studies with non-English speakers, with children, with mental nurses, etc. reveal that, although the general behavioural model of least effort is fulfilled, the above-mentioned equations in many cases do not fit observed values.[3] Due to poor fit, modifications over the decades have been proposed for the original Condon–Zipf equation.

It can be confirmed that the exponent for rank (Eq. 1) is not always equal to –1, but rather it can vary. In such cases, the expression proposed by Booth and Federowicz can be used:

$$F = \frac{k_b}{R^B} \tag{3}$$

where B is the Booth and Federowicz exponent and k_b is the Booth and Federowicz constant.

Brookes, finding that words of greatest frequency diverge while the rest present a slope equal to –1, proposed an equation that includes a parameter which modifies the rank [4]:

$$F = \frac{k_{brk}}{R + a} \tag{4}$$

where a is the Brookes parameter, and k_{brk} the Brookes constant. Thus,

$$a = \frac{k_{brk}}{1 + r_{\max}} \tag{5}$$

where $r_{\max}$ is the maximum rank of the distribution or number of different words.

However, the greatest modification (and the oldest one) of Zipf's Law is that of Mandelbrot, in a failed attempt to demonstrate the Condon–Zipf equation using lexicographic trees,[5-7] which simultaneously covers that of Brookes as well as that of Booth and Federowicz,[8,9] including therefore one parameter in the rank and one parameter in the exponent:

$$F = \frac{k_m}{(R + m)^B} \tag{6}$$

where k_m is the Mandelbrot constant and m is the Mandelbrot parameter.

Equation 6 is the fundamental equation of the model fractal proposed by Mandelbrot. This model considers that the natural space or the social space is constituted by fractional dimensions where the objects present structures self-similars like, for example, the branches of a tree.

It should be emphasized that in these equations the constants k_z, k_{brk}, k_b, and k_m are different and need not present equal values. In addition, the analysis of Eq. 6 indicates that when $B = -1$, the Brookes expression is reduced; when $m = 0$, it is identical to that of Booth and Federowicz; and logically when $B = -1$ and $m = 0$, we get the original Condon–Zipf equation.

From a practical standpoint, it is helpful to be able to determine easily the two parameters of Eq. 6 of Mandelbrot – that is, B and m. The slope B is calculated without difficulty by linear regression on the points of the straight fraction and m by the following formula:

$$m = \left(\frac{k_b}{F_m(1)} \right)^{\frac{1}{|B|}} - 1 \qquad (7)$$

In a broad review, together with the contribution of new data, Meadow et al. confirmed that the Mandelbrot equation is in general the one that best fits the 35 distributions analysed. Nevertheless, if the words belong to artificial languages, such as DIALOG or OAK II, the slopes prove strongly negative and the regression coefficients are found to be quite bad. With respect to the application to bibliographic descriptors, these authors analysed a distribution provided by Weinstock et al. and found the slope B to be close to –0.5 and the parameter m approximately 1. The regression coefficient is slightly greater than 0.8, and therefore the fit is rather mediocre.[3]

It might be asked, therefore, whether controlled artificial languages respond well to the Zipf–Mandelbrot distribution. In this sense, Ruiz-Baños analysed the descriptors of the group of articles gathered in the database Francis on archaeology from 1980 to 1993 and found substantial deviations between the frequencies or occurrences of these and the predictions by the equations Condon–Zipf, Booth and Federowicz, Brookes, and Mandelbrot.[10] This researcher also found that the distribution can be divided into three zones, each reproducible by a negative exponential term. The first zone represents the descriptors called the main ones, these being situated mainly in the centre of the network of themes shown by co-words analysis. The second zone is occupied by the so-called "thematic" descriptors – that is, those that form part of a theme directly linked to the main word. Finally, the rest of the words are the extra-thematic descriptors, which form part of the network though not within any specific theme.[10,11]

Bradford's Law

S. C. Bradford, chemist and librarian at the Science Museum of South Kensington (London), showed great interest in documentation. He published a short article[12] that can be considered one of the beginnings of bibliometric studies, in which he compiled bibliography on "lubrication, 1931-June 1933" and "applied geophysics, 1928–1931".

Also, he analysed the productivity of scientific journals, introducing his famous law, which stated in its "verbal form" the following:

> "If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as $1 : n : n^2 \ldots$"

This statement introduces for the first time the concept of "nucleus", which coincides, according to Bradford, with the first area resulting from dividing articles of a subject matter given in equal parts. To express mathematically the division of the works into areas of equal size, we could, borrowing the terminology of Egghe and Rousseau,[13-15] use the equation:

$$R(r) = iy_0 \tag{8}$$

where *R(r)* represents the accumulated articles on a given subject matter, *R* represents the accumulated journals on a given subject matter, *i* the number of Bradford zones, and y_0 productivity of the journals of the nucleus.

Continuing with the terminology of Egghe and Rousseau, we find that the number of accumulated periodical publications, *r*, is determined by the expression:

$$r = (1 + k + k^2 + k^{i-1})r_0 \tag{9}$$

where *k* is the Bradford multiplier, r_0 represents the accumulated journals of the nucleus for a given subject matter.

It should be noted that the Bradford multiplier, *k*, was represented originally by Bradford with the letter "*n*". The most usual way that Eq. 9 is found in the literature is:

$$r = r_0 + kr_0 + k^2 r_0 + \ldots + k^{i-1} r_0 \tag{10}$$

The Bradford distribution has a mixed character. According to Brookes, it consists of two parts: the first, curved, describes the nucleus, which need not coincide with the Bradford's nucleus or first part; and a second part that is logarithmic-linear.[16,17] The mathematical expression of the linear fraction proposed is:

$$R(r) = a \log\left(\frac{r}{s}\right) \tag{11}$$

where *a* is the slope of the straight fraction of the Bradford distribution, and *s* is the Brookes parameter.

For formal matters, when $s<1$, the Brookes equation in practice may not be the most appropriate for the fit of the observed values.[18] The equation proposed instead is:

$$R(r) = a \log\left(\frac{r}{x_1}\right) + y_1 \tag{12}$$

where x_1, y_1 are the coordinates of any point of the straight fraction of the Bradford distribution.

Much more simple and direct is to fit the straight fraction to an equation of the type:

$$R(r) = a \log r + c \tag{13}$$

where c is the ordinate at the origin.

This equation has been applied with very good results for the evaluation of nucleus of journals that publish the works produced by the University of Granada.[19]

Eq. 13 is a more easily handled than the Brookes equation (Eq. 11). They are, of course, identical with

$$c = -a \log s \tag{14}$$

The expressions above (Eqs 11, 12 and 13), which can be called the Brookes–Ferreiro–Bradford Law, fit the experimental values corresponding only to a straight fraction. The fit to the straight fraction and to the nucleus can be achieved by Leimkuhler's equation, although with this expression Groos' droop cannot be fit, as will be seen below:[20]

$$R(r) = a \log(1 + br) \tag{15}$$

where b is Leimkuhler's parameter.

This expression was considered by its author as exact,[21] a qualifier that has been criticized as being excessive.[22]

When r is very large the term br is far greater than unity, so that we can ignore 1 of the logarithm:

$$R(r) = a \log(br) \qquad \text{for a high } r \tag{16}$$

If we have:

$$b = \frac{1}{s} \tag{17}$$

the equation is consistent with the equation of the straight fraction of Brookes and therefore equivalent to that of Ferreiro and Eq. 13.

Leimkuhler's equation expresses the exact same verbal statement as Bradford's.[13] It can be demonstrated that:

$$a = \frac{y_0}{\log k} \tag{18}$$

$$b = \frac{k - 1}{r_0} \tag{19}$$

These latter formulas provide an easy way to determine the Leimkuhler parameters (a and b) from the observed values. We should take into account that k is determined by:[14]

$$k = \left(e^{\gamma} y_m\right)^{\frac{1}{i}} \tag{20}$$

where e = 2.71828, γ = 0.5772 (Euler's number), and y_m represents the articles of the most productive journal.

Roughly, k is equal to:

$$k = \left(1.781 y_m\right)^{\frac{1}{i}} \tag{21}$$

Frequently, we find not only that the Bradford distribution presents an initial area or nucleus and later a straight fraction, but we may also find an area, beyond the straight line, in which the number of articles slowly increases. This new curvature is called the Groos droop.[23] Figure 1 presents the distribution of the three fractions, including that of Groos.

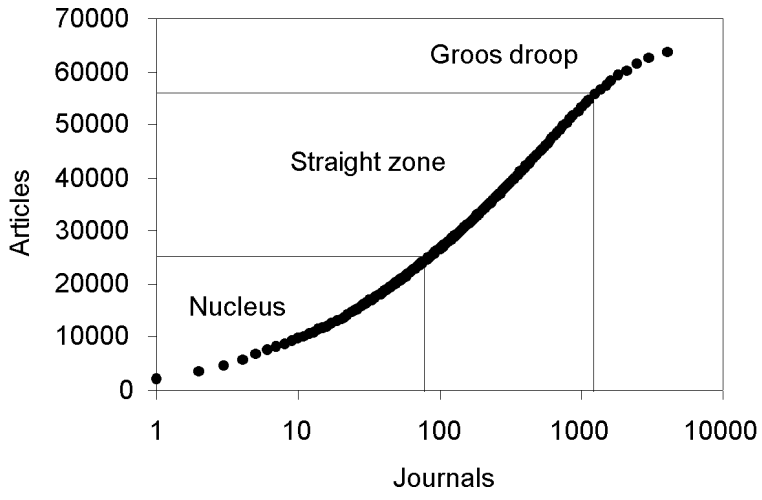

Figure 1. Fractions or zones of the Bradford distribution.
Data: Journals published between 1993 and 2002 on surfactants, cosmetics, and fragrances

Neither of the equations proposed above, that of Brookes–Ferreiro or that of Leimkuhler, take into account the possibility of this droop. A fit of the entire distribution that includes the three fractions could be made using an equation of the Lotka type with an exponent greater than 2. The fit of the observed values is in general deficient.[24] The inflexion point can be determined in a simple way analogous to the determination of the nucleus using the Brookes–Ferreiro equation. We consider the nucleus to be those values that separate a certain percentage of the Brookes–Ferreiro

equation before reaching the straight zone (more than 1% or 2%) and the Groos droop those that also separate a given percentage, but after the straight zone.

None of the above expressions fit the entire Bradford distribution: simultaneously nucleus, straight fraction, and Groos droop. Rousseau had deduced an equation based on the parameters of a fit by Lotka, which is called generalized Leimkuhler and which enables a fit of the entire ranking.[25]

$$R(r) = \frac{C}{2-\alpha}\left[y_m^{2-\alpha} - \left(y_m^{1-\alpha} - \frac{\alpha - 1}{C} r \right)^{\frac{2-\alpha}{1-\alpha}} \right] \tag{22}$$

where C and α are two parameters to be evaluated, y_m is the number of articles of the most productive journal, r is the number of accumulated journals, and $R(r)$ is the number of accumulated articles.

At least in some examples offered by Egghe and Rousseau, the fit achieved is quite acceptable. The disadvantages that can be appreciated in this expression include a certain complexity, the impossibility of being linear to perform a simple fit by linear regression, and the need to make an indirect evaluation algorithm from the parameters.

Lotka's Law

Let us consider a set of authors that publish on a given subject over a rather long time period. If we arrange the authors according to their productivity, we find that the immense majority publish few works, while only a select portion are highly productive. The first expression that related the number of authors to their productivity, given by Lotka, indicates that the number of authors that publish a certain quantity of works is inversely proportional to the square of these works:[26]

$$A(\mathrm{R}) = \frac{A(1)}{\mathrm{R}^2} \tag{23}$$

where $A(\mathrm{R})$ is the number of authors that publish R works, R is the number of works that an author publishes, and $A(1)$ is the number of authors that publish only one work.

Subsequent studies in different subject areas have confirmed the accuracy of the above inverse power expression, although with the exception that the exponent is not always two but rather a variable value. Consequently, Lotka's Law is generalized by the following equation:

$$A(\mathrm{R}) = \frac{A(1)}{\mathrm{R}^m} \tag{24}$$

where m is the Lotka exponent.

The value of the Lotka exponent is related to productivity by a scientific community. It depends, moreover, on the subject area considered, on the community of scientists studied, and even, when maintaining the above variables constant, on the historical moment.[10]

Lotka's Law is analogous to Pareto's Law of the distribution of income, although in this latter distribution the exponent tends rather towards 1.5 than towards 2.0. The final reason for this analogy is that in both cases an increased effort is followed by a logarithmic increase in its results, as often occurs with human stimuli, as indicated in the Fechner or Weber Laws in Experimental Psychology.[27]

Lotka's Law presents good fits of observed values in the area of low production of works. On the other hand, when we approach the points of very productive authors, the fits by regression substantially worsen. Therefore, the value of $A(1)$ is usually prone to considerable error. For a better distribution, a calculation method that improves the results has been proposed.[28] The algorithm consists of submitting the logarithms of the authors and works to a linear regression in the usual way. From the slope, the Lotka parameter is determined and an improved $A(1)$, using decimals for the percents, by the following equation:

$$A(1) = \frac{1}{\sum_{R=1}^{P-1} \frac{1}{R^m} + \frac{1}{(m-1)P^{m-1}} + \frac{1}{P^m} + \frac{m}{24(P-1)^{m+1}}} \tag{25}$$

where P is an arbitrary value greater than or equal to 20.

It has been demonstrated theoretically that in the particular cases in which $m = 2$ and $m = 4$, using decimals for the percents, the value of $A(1)$ is[14]:

$$\begin{aligned} m = 2 &\rightarrow A(1) = \frac{6}{\pi^2} \\ m = 4 &\rightarrow A(1) = \frac{90}{\pi^4} \end{aligned} \tag{26}$$

If the Pao equation is used for these values of m, using $P = 20$, the differences between the equation and the theoretic values are less than 1/110,000 and 1/25,000,000, respectively.

Objectives

Since their advent, the bibliometric laws of Zipf, Bradford, and Lotka have been evolving to achieve a fit of the empirical values. Depending on the particular case, better fits are attained with simpler or more complex expressions. That is, though Joyce's *Ulysses* can be fit adequately with the Condon–Zipf equation (the simpler

expression), the artificial language of DIALOG cannot be fit, even using the more complex Zipf–Mandelbrot equation.

The aim of the present paper, part of a series of papers related to the development of a Unified Scientometrics Model, is to demonstrate that, in general, the bibliometric laws of Zipf, Bradford, and Lotka, being equivalents of each other, do not fit the empirical values because the underlying fractal model is not sufficient. Furthermore, the analysis is restricted not only to frequency of words, authors and journals, but also includes econometric and demographic aspects.

Materials and methods

The scientific field analysed involves surfactants and related materials. The CoPalRed© system has been used on a set of 63,543 bibliographical references of scientific articles published between 1993 and 2002. The query used is:

"SURFACTANT* OR DETERGENT* OR TENSIDE* OR CLEANER* OR LAUNDRY* OR FRAGRANCE* OR PERFUME* OR FLAVOR* OR ODOR* OR (ESSENTIAL SAME OIL) OR COSMETIC* OR TOILETRY* OR SOAP*"

Activity was studied by countries, by institutions and research laboratories, by researchers, by journals, and even by frequency of the terms used as key words in documents. Also, correlations were made with the economic resources available and the total population.

Production of articles by country and its relationship with the GNP and the population

The world distribution by country of any tangible goods, such as gross national product (GNP), automobile manufacturing, sulphuric-acid production, or even the inhabitants of these countries, or any other entity, is usually approached using Pareto's Law (Lotka's Law of Scientometrics), or simply by applying Rule 80/20 (simplified version of Pareto, which states that 80% of the production is belongs to 20% of the entities). Therefore, following the tradition, the production per country of articles on surfactants and related fields can be analysed by fitting the values to the Lotka–Pareto Law.

Figure 2 provides a representation of the Lotka–Pareto type. The fit is quite deficient, as the regression line can hardly be adapted to such an extreme dispersion (the coefficient of determination, R^2, barely reaches the very low value of 0.4). Also, the area of high production gives the regression line a lower slope than needed for the low T zone. By Pao approximation, the ordinate at the origin can be improved (Eq. 25), although, due to the great scattering of the cloud of points, the result can never be

satisfactory.[28] The conclusion is evident: the production of articles by country cannot be fit by the Lotka–Pareto Law. For the equivalence of the bibliometric laws, a fit would not be provided by Zipf's Law (in any of its versions) or Bradford's (except, perhaps, partial fits of only a straight fraction of the distribution).

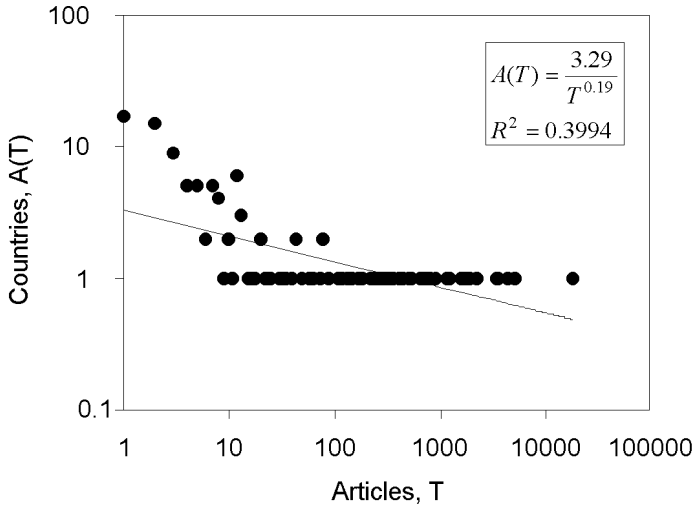

Figure 2. Production of articles by country. Fit to the Lotka–Pareto Law

Nevertheless, as confirmation, we have represented this production against rank according to a Zipf-type distribution. Specifically, we used the Zipf–Mandelbrot Law, as it is the most flexible of all of the laws (Eq. 6). The mathematical treatment is executed as indicated above. First, consideration is given to the highest ranking points, which in principle should be aligned in a double-logarithmic diagram and adjusted by regression to a function of the Booth–Federowicz type (straight line of the figure, with $R^2 = 0.993$). From this fit, the B parameter and the k_b are determined. In addition, $F_m(1)$ is considered to be the frequency of the most productive country. These values are replaced in Eq. 7. The value of m found is used in the Mandelbrot equation (Eq. 6), causing K_m to be identified with k_b determined by the foregoing regression. The final result is the curve in Figure 3, the equation of which is:

$$F(r) = \frac{2 \times 10^8}{(r + 10.54)^{3.79}} \tag{27}$$

The fit proves better than with that of the Lotka–Pareto equation, but is still not completely satisfactory. Figure 4 presents the values calculated against the observed

ones, reflecting that a linear regression gives a slope of 24% greater than unity and a poor determination coefficient.

$$F(r)_{calc.} = 1.2396 F(r)_{Obs.} \quad R^2 = 0.840 \tag{28}$$

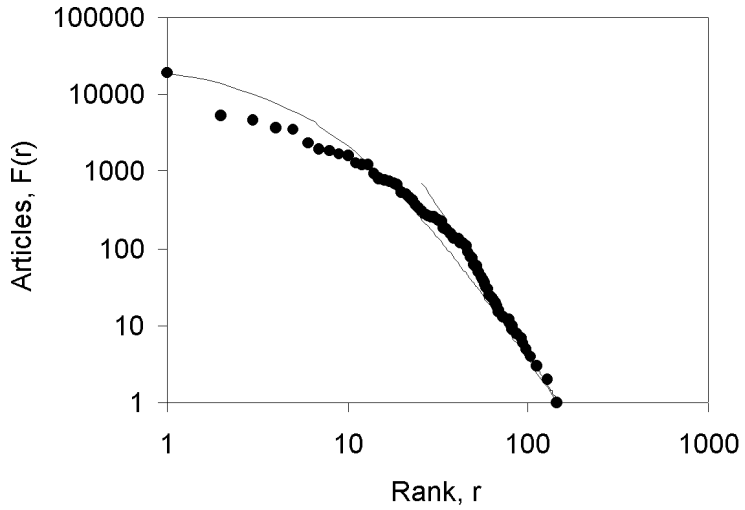

Figure 3. Production of articles by country. Fit to the Zipf–Mandelbrot Law

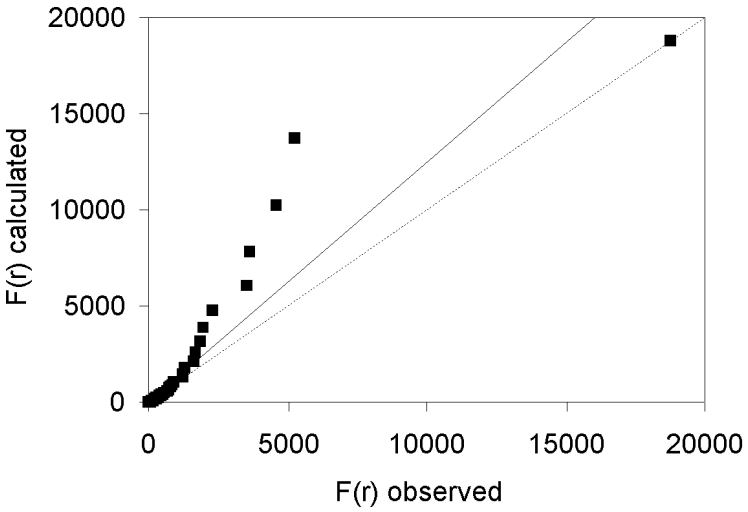

Figure 4. Production of articles calculated, F(r) calc, vs. observed

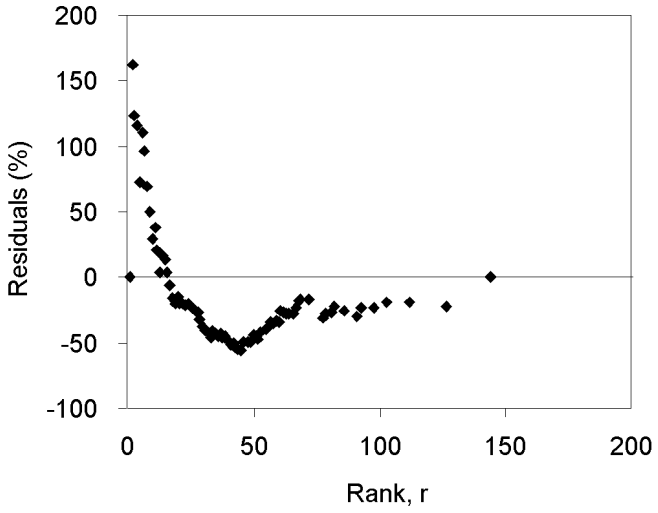

Figure 5. Production of articles by country. Analysis of residuals for the Zipf–Mandelbrot Law

The analysis of the results (Figure 5) clearly reflects a non-random distribution, but rather a definite pattern (indicated by the broken line). Furthermore, these residuals in many cases present values higher than 50% (in some cases higher than 100%!). In short, the production of articles need be fit to an another, as yet unknown, equation – not the Law of Zipf–Mandelbrot, but one having a certain similarity to these.

This point raises the following question: Does the production of articles on surfactants, by country, not adjust to Lotka, Pareto, or Zipf–Mandelbrot Laws because it is a unique or anomalous case, as in the example of the GNP or the number of inhabitants, which, according to any manual on economy or statistics in use, should completely fit the Law of Pareto and Zipf–Mandelbrot?

Of the 144 countries that have produced articles on surfactants, the GNP was found for the year 2000 for 127 of them.[29-31] For this case, which is considered the most general of all and a paradigm, the same lack of fit would imply that established socio-economic questions should be reviewed and would confirm the results regarding the production of articles on surfactants.

The same treatment was made as in the preceding case, fitting the values to a distribution of the Zipf–Mandelbrot type, the most general and flexible of all. The straight fraction used goes from the point of rank 40 to rank 119 (those above 119 disappeared because they fell too abruptly). The resulting equation is:

$$F(r) = \frac{6 \times 10^6}{r^{3.01}} \quad R^2 = 0.9935 \tag{29}$$

From this point, we determined the Mandelbrot parameter, *m*, for which the resulting Zipf–Mandelbrot Law is:

$$F(r) = \frac{6 \times 10^6}{(r + 7.41)^{3.01}} \quad R^2 = 0.871 \tag{30}$$

With the GNP, exactly the same occurs as with the production of articles on surfactants: it does not satisfactorily obey the Zipf–Mandelbrot Law. Moreover, conformance to the Zipf–Mandelbrot Law or Pareto's Law would imply the accuracy of the well-known Rule 80/20 that literally states:

80% of world wealth is produced by 20% of the countries.

If we take into account the data that we have used (from UNESCO and the International Monetary Fund), we find that:

Really, 80% of world wealth is produced by only 9% of the countries.

This has two implications:

1. The laws of Lotka–Pareto and their equivalents (Zipf and Bradford) must be revised.
2. The world is substantially poorer than usually purported…

In addition, if we analyse the production of scientific articles on surfactants and related fields reported by the SCI, we find also that: 80% of the production of articles on surfactants are produced by only 9% of the countries.

The case of the surfactants is not, as can be seen, a particular exception, but rather conforms to general behaviour. Furthermore, the capacity of research is proportional to the economic resources available to the researchers involved.

Another fundamental element in all the socio-economic approaches, and scientometric ones is population distribution. Also, it is usually claimed that population obeys the Zipf–Mandelbrot Law (see for example Ref. 6). This assertion should be tested.

Figure 6 shows the fit of world population to Zipf–Mandelbrot. Again, we find that this law and its equivalents, such as Lotka–Pareto do not hold. In fact, the resulting mathematical expression, using as the straight zone the part that goes from rank 50 to 100 (although we could have taken some other area, as the distribution is very curved), is as follows:

$$F(r) = \frac{232180}{(r + 7.41)^{2.41}} \quad R^2 = 0.8328 \tag{31}$$

Two final questions remain: What is the distribution of the population of the articles per capita and the distribution of the production per economic unit? It might be suspected that if we divided a distribution similar (but not equal) to the Law of Zipf–Mandelbrot by another one also similar to Zipf–Mandelbrot, the resulting distribution could reveal the factor that makes it impossible for the two distributions to make an exact fit to Zipf–Mandelbrot.

Figure 7 shows a good fit in the log-linear diagram, indicating that the function is an exponential type. The regression offers the following equation:

$$Pc(r) = 113.42e^{-0.057r} \quad R^2 = 0.990 \tag{32}$$

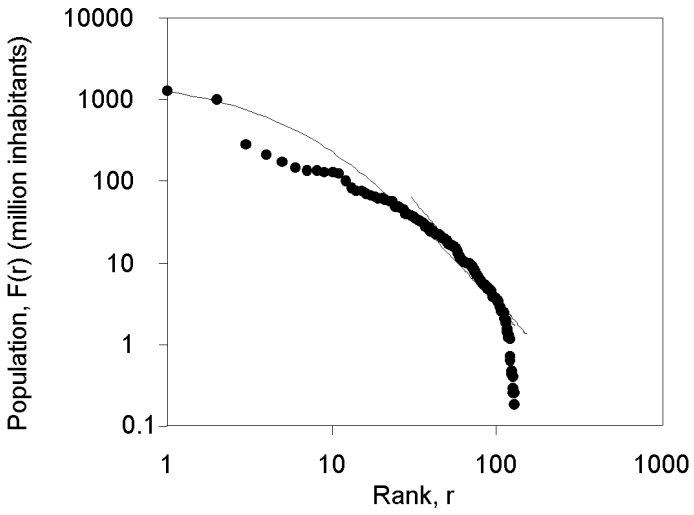

Figure 6. Distribution of population by country. Fit to the Zipf–Mandelbrot Law

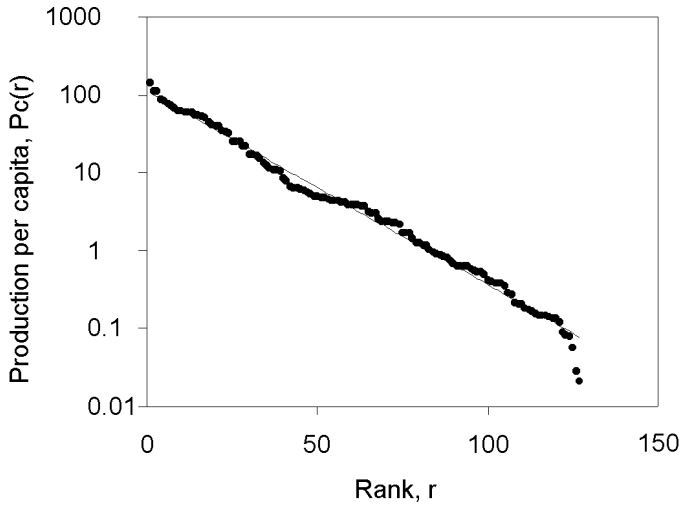

Figure 7. Production per-capita of articles on surfactants

A similar result is reached in the production of articles per billion dollars of GNP: an exponential equation (Figure 8)

$$P_{PIB}(r) = 8.453e^{-0.032r} \quad R^2 = 0.946 \tag{33}$$

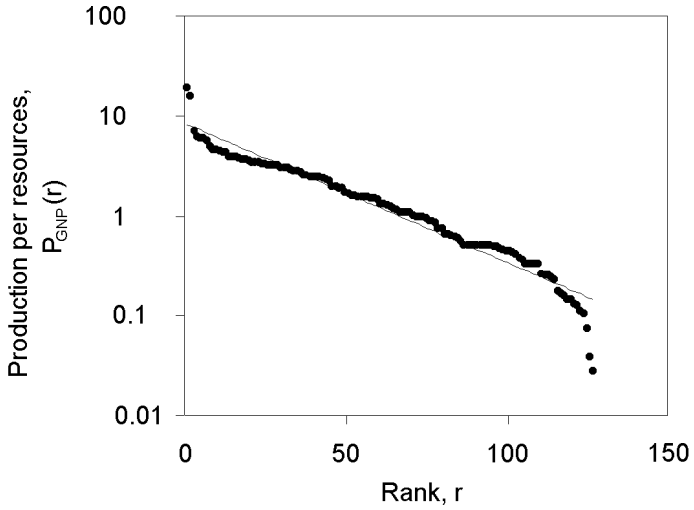

Figure 8. Production of articles on surfactants per billion dollars GNP

All this implies that the underlying model is a hybrid model between an inverse power function (such as that of Zipf–Mandelbrot) and a negative exponential function.

An illuminating qualitative confirmation can be made. According to Figure 9, if we resort to the use of an inverse power representation (log-log diagram), we cannot align the points, which remain in a convex curve.

If we use a negative exponential representation (log-linear diagram), we cannot align the points, either, which remain in the form of a concave curve. Figures a) and b) form mirror images. The conclusion is clear: the model that we seek should undoubtedly be a hybrid between an inverse power function and a negative exponential function.

To reaffirm even further whether the proposal is valid, it is necessary to test whether the situation posed for the case of production of articles by country occurs in a similar way, to a greater or lesser degree, for the production of laboratories, journals, authors, and above all for the distribution of descriptors (typical case of Zipf's Law).

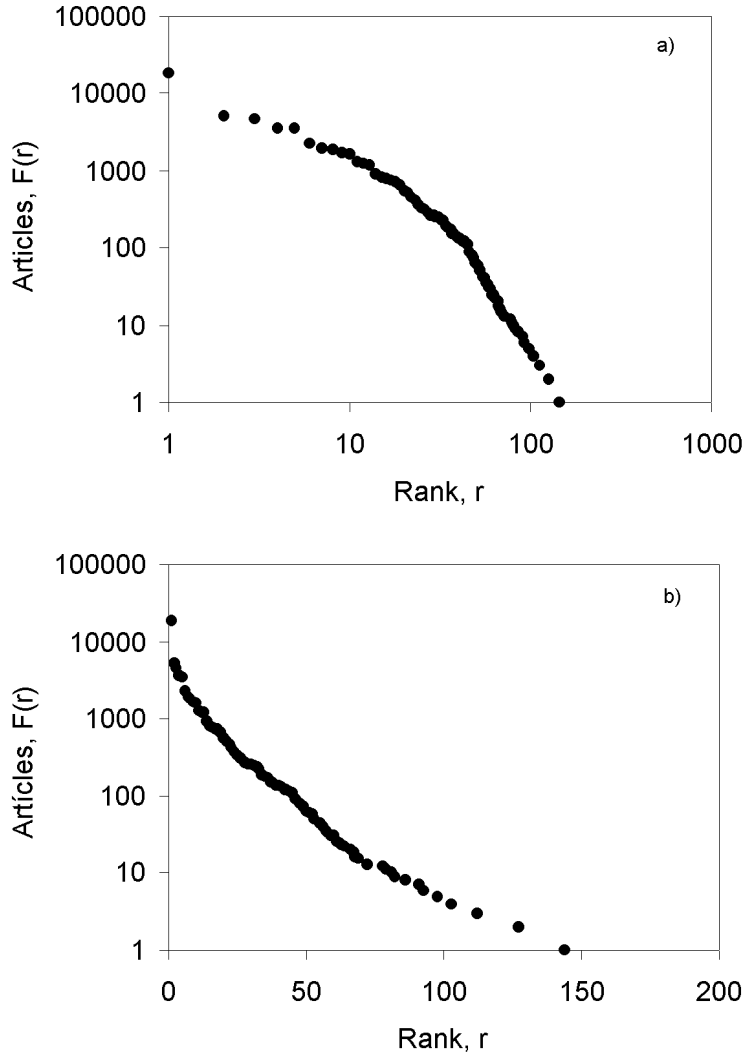

Figure 9. Production of articles by country. Comparison between inverse power representation and negative exponential representation
a) Inverse power representation; b) Negative exponential representation

Production of articles per laboratory, journal, and authors. Analysis of the distribution of descriptors

Figure 10 confirms that the distribution pattern, in all cases, is similar. Regardless of the type of actor considered, the production of items (articles or descriptors) is governed by a common model. This is, therefore, the empirical support of the model of Information Production Processes (IPP) proposed by Egghe and Rousseau. In addition, as demonstrated by these authors, this similarity in behaviour would have been exactly the same if the representation were of the frequency type vs. rank (Zipf's Law, as in this case), accumulated frequency vs. rank (Bradford's Law), or number of sources or actors that produce a given number of items (Lotka–Pareto Law).[14]

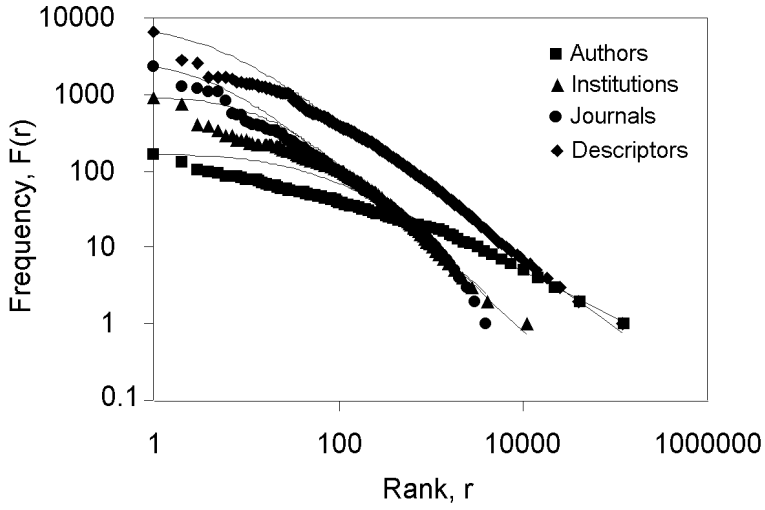

Figure 10. Production of articles by author, institution, and journal. Distribution of frequency and descriptors. Fit by the Zipf–Mandelbrot Law

The solid lines represent the best fit, according to Zipf–Mandelbrot, that can be achieved. It is shown that, although in general terms this equation corresponds to the distribution profile, the fit is deficient, as corroborated in the previous section concerning the production of articles by country. Table 1 presents the parameters of the fit, indicating the coordinates of the initial and final points that have been taken as the straight zone of the distribution, the k_m constant, the B exponent, the m constant, and the determination coefficient R^2. It can be seen that the distribution of journals made the best fit (even better than that of the descriptors, in principle being the favourite for the

best result) while that of the institutions and laboratories made the worst. In no case did the Zipf–Mandelbrot Law adequately represent the phenomena under study for any of the actors considered.

Figure 11 offers the comparison between the calculated and observed values. In general, the errors accumulate for high production/frequency, with all the cases following a similar pattern.

Table 1. Parameters of fit for the overall distributions to the Zipf–Mandelbrot Law

	Authors	Institutions	Journals	Descriptors	Countries
Initial point of the straight zone	1019	501	101	6673	47
Final point of the straight zone	125730	11229	4071	119308	144
k_m	1323	27332	12700	27565	$2x10^8$
B	0.608	1.130	1.026	0.900	3.79
m	29.47	19.57	4.33	4.06	10.54
R^2	0.891	0.825	0.944	0.883	0.840

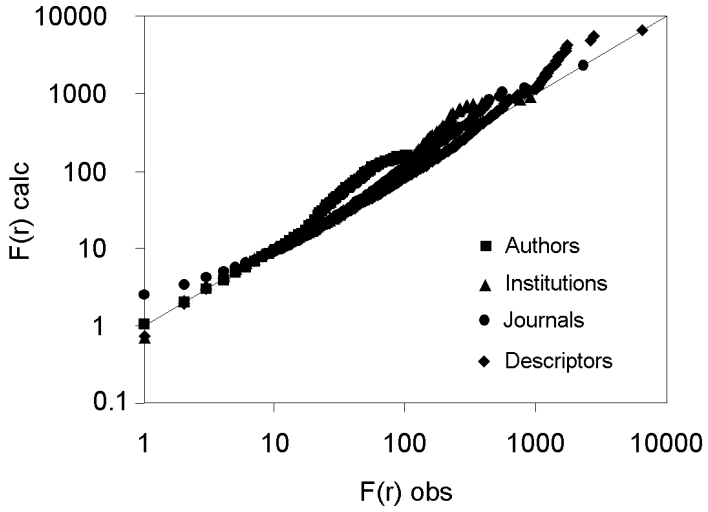

Figure 11. Observed vs. calculated values in the overall distributions of authors, institutions, journals, and descriptors for the Zipf–Mandelbrot fit

The analysis of the residuals corroborates even more clearly the existence of a common pattern. In no case were the residuals distributed at random, indicating that the Zipf–Mandelbrot Law is inadequate, because of an underlying model that provides a better explanation for the empirical values (Figure 12).

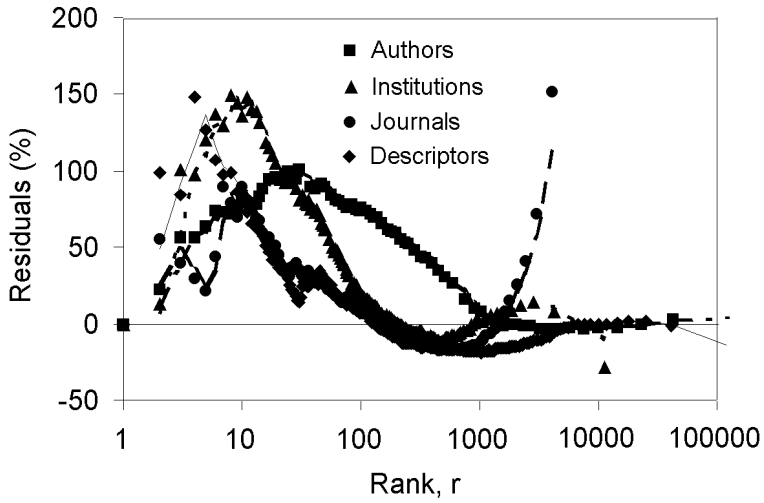

Figure 12. Residuals for Zipf–Mandelbrot fits for authors, institutions, journals, and descriptors

Conclusions

The above results clearly lead to the conclusion that it is vital to find a model that adequately fits the empirical values. This implies a revision of the fundamentals of the fractal model, since, as shown, the inverse power equations (of the Zipf–Mandelbrot type) are not adequate, as they need to include exponential terms.

These modifications not only affect Bibliometrics and Scientometrics, but also, for the generality of the fractal model, apply to Economy, Demography, and even Natural Sciences in general.

*

This study was supported by the Spanish Ministry of Science and Technology through projects no. 1FD97-0931 and PB1998-1293.

References

1. CONDON, E. U., Statistics of vocabulary. *Science*, 68 (1928) 1733.
2. ZIPF, G. K., *Human Behaviour and the Principle of Least Effort*, Adisson-Wesley Press, Inc, Cambridge, 1949.
3. MEADOW, C. T., WANG, J., STAMBOULIE, M., An analysis of Zipf–Mandelbrot language meausures and their application to artificial languages. *Journal of Information Science*, 19 (1993) 247–258.
4. BROOKES, B. C., Ranking techniques and the empirical log law. *Information Processing & Management*, 20 (1984) 16–37.

5. MANDELBROT, B. B., An informational theory of the stadistical structure of language. In: W. JACKSON (Ed.) *Communication Theory*, pp. 486–502. London, Butterworths Scientific Publications, 1953.
6. MANDELBROT, B. B., *The Fractal Geometry of Nature*, Freeman, New York, 1977.
7. MANDELBROT, B. B., Structure formelle des textes et communication (deux études). *Word*, 11 (1954) 424.
8. FEDEROWICZ, J. E., A zipfian model on automatic bibliographic system: an application to MEDLINE. *Journal of the American Society for Information Science*, 33 (1982) 223–232.
9. FEDEROWICZ, J. E., The theorical foundation of Zipf's law and its application to the bibliographic database environment. *Journal of the American Society for Information Science*, 33 (1982) 285–293.
10. RUIZ-BAÑOS, R., *Ciencimetría de redes. Análisis de la investigación internacional sobre Arqueología mediante el Método de las Palabras Asociadas (1980-1993).* Ph. D. Thesis. Universidad, Granada, 1997.
11. RUIZ-BAÑOS, R., BAILÓN-MORENO, R., JIMÉNEZ-CONTRERAS, E., COURTIAL, J. P., Structure and dynamics of scientific networks. Part 2: The new Zipf's Law, the cocitations's clusters and the model of the presence of key-words. *Scientometrics*, 44 (1999) 235–265.
12. BRADFORD, S. C., Sources of informations on specific subjects. *Engineering*, 137 (1934) 85–86.
13. EGGHE, L., Consequences of Lotka's law for the law of Bradford. *Journal of Documentation*, 41 (1985) 173–189.
14. EGGHE, L., ROUSSEAU, R., *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam, etc., 1990.
15. ROUSSEAU, R., The nuclear zone of a Leimkuler curve. *Journal of Documentation*, 43 (1987) 322–333.
16. BROOKES, B. C., The Bradford law: a new calculus for the social sciences? *Journal of the American Society for Information Science*, July (1979), 30 (4) 233–234.
17. BROOKES, B. C., Bradford's law and the bibliography of science. *Nature*, 224 (1969) 653–656.
18. FERREIRO-ALAEZ, L., MENDEZ, A., Linealidad de las dispersiones Bradford. *Revista Española de Documentación Científica*, 3 (1980) 201–211.
19. JIMÉNEZ-CONTRERAS EVARISTO, *Difusión de la literatura científica granadina reciente* (1975-87). Granada, Universidad de Granada,1993.
20. LEIMKUHLER, F. F., The Bradford distribution. *Journal of Documentation*, 23 (1967) 197–207.
21. LEIMKUHLER, F. F., An exact formulation of Bradford's law. *Journal of Documentation*, 36 (1980) 285–292.
22. BROOKES, B. C., A critical commentary on Leimkuhler's 'exact' formulation of the Bradford law. *Journal of Documentation*, 37 (2) (1981) 77–88.
23. GROOS, O. V., Bradford's law and the Keenan-Atherton data. *American Documentation*, 18 (1967) 46.
24. EGGHE, L., The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, 16 (1990) 17–27.
25. ROUSSEAU, R., Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information Studies*, 25 (1988) 150–178.
26. LOTKA, A. J., The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16 (1926) 317–323.
27. PRICE D. J. D. S., *Little Science, Big Science*, Columbia Univ. Pt., New York, 1963.
28. PAO, M. L., Lotka's law: a testing procedure. *Information Processing & Management*, 21 (1985) 305–320.
29. ANZIL, F. (2003), Ecolink.com. Retrieved from http://www.econlink.com.ar/datos/mundo/pbipercapita.shtml
30. UNESCO (2003) Science and Technology: UNESCO UIS. Retrieved from http://portal.unesco.org/uis/TEMPLATE/html/sc_consult.html
31. INTERNATIONAL MONETARY FUND (2003), International Monetary Fund Home Page. Retrieved from http://www.imf.org/