Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 155–160
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

On the behavior of journal impact factor rank-order distribution

R. Mansilla [a], E. Köppen [a], G. Cocho [b], P. Miramontes [c,*]

[a] *Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*
[b] *Instituto de Física, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*
[c] *Facultad de Ciencias, Universidad Nacional Autónoma de México, México 04510, D.F., Mexico*

Received 25 September 2006; received in revised form 30 December 2006; accepted 4 January 2007

Abstract

An empirical law for the rank-order behavior of journal impact factors is found. Using an extensive data base on impact factors including journals on education, agrosciences, geosciences, mathematics, chemistry, medicine, engineering, physics, biosciences and environmental, computer and material sciences, we have found extremely good fittings outperforming other rank-order models. Based in our results, we propose a two-exponent Lotkaian Informetrics. Some extensions to other areas of knowledge are discussed.

Keywords: Zipf's law; Lotkaian Informetrics; Power laws; Impact factors

1. Introduction

Quantitative studies in linguistics have a long lineage. Due to the extreme complexity of languages, these studies have been mainly based on statistical properties of words in literary corpora. Outstanding early examples of these studies are Estoup (1916), Dewey (1923) and Condon (1928). However, the most influential contribution on this topic is by Zipf (1949). In his work it appears what is today known as Zipf's law which can be formulated as follows: let $f(r)$, $r = 1, \ldots, N$, be the relative frequency of the words (the number of times a word appears divided by the total amount of words) in a text in decreasing order. Then Zipf's law states that:

$$f(r) = \frac{K}{r^{\alpha}}. \tag{1}$$

In this case, the items are words taken from a given text, the most abundant word takes the first place ($r = 1$), the second one takes the following place ($r = 2$) and so on. The fact that the mathematical expression of the law is a negative exponent power law implies that the law is a straight line with negative slope α when plotted in log–log scales. K is a proportionality constant with no phenomenological interest. This empirical law has found applications in a wide range of natural and human phenomena (Li, 2003). The case when $\alpha \simeq 1$ is of particular interest because it implies self-similarity.

* Corresponding author. Fax: +52 55 5622 4859.
E-mail address: pmv@fciencias.unam.mx (P. Miramontes).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2007.01.001

The exact mechanism behind Zipf's law still remains a mystery so far. However, it is important to remark that the presence of power laws implies in general that the underlying mechanism is neither stochastic or regular. Power laws are the signature of correlated (colored) noise possibly indicating an "edge of chaos" dynamics (Langton, 1990) and all the rich phenomena associated (McElvey, 2001) or could be as well a clue to self-organized criticality (Bak, Tang, & Wiesenfeld, 1987).

The main drawback of Zipf's law was the bad fitting at very high and very low frequencies in the word counting problem. An improvement over the Zipf's law was proposed by Mandelbrot (1954):

$$f(r) = \left[\frac{N+\rho}{r+\rho}\right]^{1+\epsilon}, \tag{2}$$

where N is the number of different words in the text and ρ, ϵ are parameters to be adjusted.

Zipf's law is a special case of Mandelbrot's. This fact, along with a complete discussion of the role of power laws in the field of Informetrics can be found in Egghe (2005).

Recently it has been reported (Le Quan, Sicilia-García, Ming, & Smith, 2002) that what Zipf found is valid for small corpora (for the size of the texts that were analyzable at that time), and that today that the computer allows the analysis of huge texts, the log–log plot shows a clear downwards bending tail instead of the predicted straight line.

Scientific productivity is another topic where the first studies date back almost a century with the works of Dresden (1922) and Lotka (1926). The law of Lotka has the same mathematical form of Eq. (1) but he already introduced bibliometric variables by using r as contributors or authors of a given paper and $f(r)$ as articles or papers themselves. Since Lotka, it is common to call "sources" the independent variable and "item" the dependent one. This way, Lotka's law states that the number of items is a power law of the sources. The branch of Informetrics related to the study of power laws is called Lotkaian Informetrics (Egghe, 2005).

Informetrics mainly deals with the relationships between sources and items. It is normal to find the pairs authors–journals or journals–bibliographies as sources and items. In this paper we explore the possibility of extending the Lotkaian Informetrics to the realm of journal impact factors (JIFs). We show as well that the rank-order JIFs plots deviate from a traditional Lotkaian equation and propose an extension to what it could be called two-exponent Lotkaian laws.

2. Impact factors

Impact factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period (Garfield, 1994), it is calculated "by dividing the number of times a journal has been cited by the number of articles it has published during some specific period of time. The journal impact factor will thus reflect an average citation rate per published article" (Garfield, 1955). The impact factor of journals is an attempt to evaluate the knowledge production published among different journals of a given field. Mainly covered by the Science Citation Index database, it is published annually since 1975 in the Journal Citation Reports.

JIFs has been the target of many criticisms (Soegler, 1997; Fröhlich, 1996) and there is a debate about its usage as a tool to evaluate research. Even the influential journal Nature states that the JIFs figures should be handled carefully (Nature, 2005). Regardless its pros and cons, the fact is that it is an every day measure of the importance of a journal and it is worldwide used (de Marchi & Rocchi, 2001).

While keeping a skeptical attitude towards the use of the JIFs to evaluate scientific research, it should be recognized that it is an outcome of the process of publication and it has became by itself a subject of scientific study.

Rank-order distribution of JIFs attracted the attention of Lavalette who (mentioned in Popescu, 2003) proposed the following law:

$$f(r) = K\left[\frac{N+1-r}{r}\right]^{b} \tag{3}$$

where N is the number of journals, r the ranking number, $f(r)$ the impact factor, and b is a parameter to be fitted.

In the next section we propose a law that outperforms Lavalette's (see Section 5).

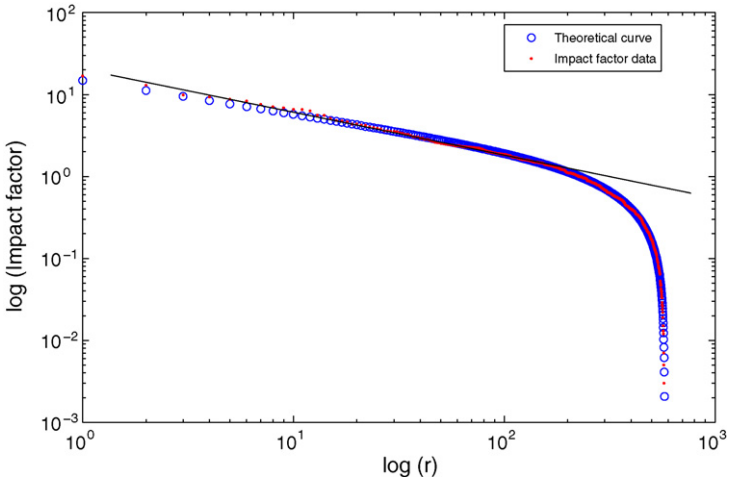

Fig. 1. Log–log rank-order plot of the impact factor data for physics journals. Notice the drop of the tail of the curve (see text).

Table 1

Scientific field	k	b	a	R^2
Physics	0.0273	0.991	0.4058	0.9999
Mathematics	0.0437	0.676	0.2622	0.9999
Computer science	0.0066	1.0626	0.2840	0.9999
Agroscience	0.0070	0.9597	0.2210	0.9999
Environmental science	0.0358	0.9357	0.2781	0.9800
Biosciences	0.0304	1.0161	0.5140	0.9999
Chemistry	0.0549	0.9733	0.4560	0.9999
Engineering	0.0033	1.0472	0.3522	0.9999
Geosciences	0.0463	0.8739	0.3505	0.9999
Material science	0.0408	0.9072	0.4477	0.9999
Medicine	0.0819	0.7735	0.4307	0.9999
Education	0.0819	0.7735	0.4307	0.9999

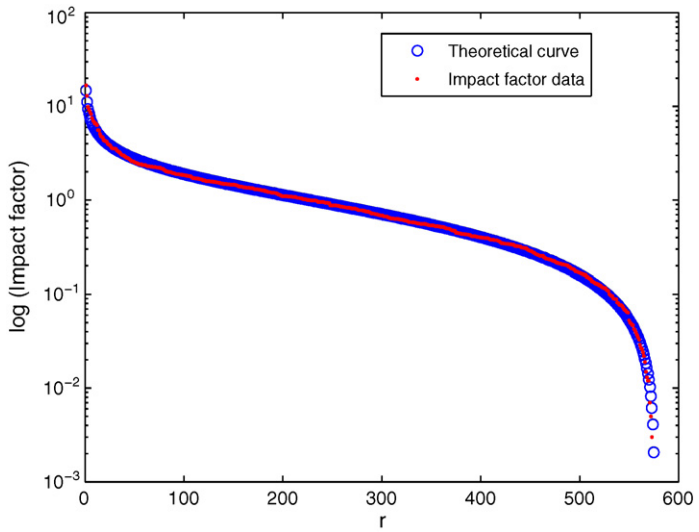

Fig. 2. Semi-log impact factor rank-order distribution for physics journals. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

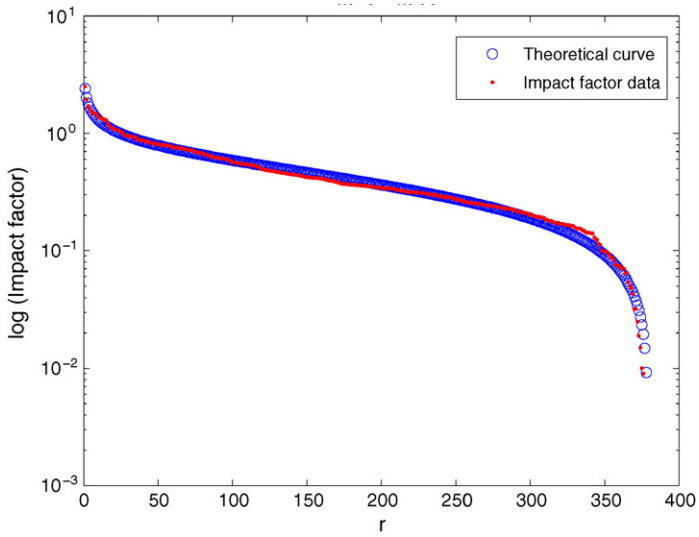

Fig. 3. Semi-log impact factor rank-order distribution for mathematics journals. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

3. Analytical expression of the law

Fig. 1 shows the log–log plot of the IF of a randomly taken field from Popescu's database (2003).

It is evident that it is not a power law (the straight line was drawn as a reference) because of the bending tail in the right side of the plot. This fact motivated us to propose a beta-like function:

$$f(r) = K\frac{(N+1-r)^b}{r^a} \tag{4}$$

$f(r)$, $r = 1\ldots, N$ represents the rank-order impact factors; K, a and b are three parameters to fit. K is a meaningless scaling factor. Notice that when $b = 0$ this equation becomes Lotka's law.

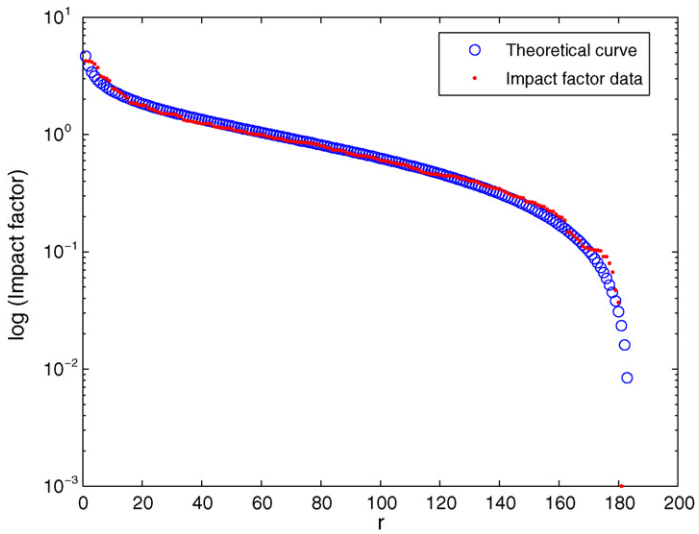

Fig. 4. Semi-log impact factor rank-order distribution for environmental sciences. Solid circles represent raw data. Hollow circles are the data evaluated in Eq. (4).

4. Results

For every set of data, we find the parameters values using the linear least squares method after transforming the coordinates to the logarithmic variable:

$$\log(f(r)) = \log(K) + b\,\log(N + 1 - r) - a\,\log(r) \tag{5}$$

Table 1 shows the values of K, b and a, as well as the coefficient of regression r^2 for impact factors of 12 disciplines. In Figs. 2–4 , the impact factors data as well as our theoretical curve for the fields of physical, mathematical and environmental sciences are shown. We used semilog plots because they are more natural when the abscissa is a rank-order variable.

The quality of the fitting is remarkable. Please notice from the analysis of Eq. (4) that the parameter a is more influential for small values of r. This fact means that for low values of r the phenomenon is nearly Lotkaian but this property is lost as the abscissae increase.

5. Concluding remarks

We have shown the excellent agreement of the data with our model. The quality of the fitting is superior to the proposal of Lavalette. From the comparison of Eqs. (3) and (4), it follows that Lavalette's law is a particular case of ours when $a = b$. Unfortunately, it is not possible to discuss the rationale behind Lavalette's law because the original paper is not available and all we know about it is a mention in Popescu's paper (2003).

The underlying proposed mechanism yielding the above-discussed behaviors often assumes a kind of "biological evolution form". For instance, Yule (1924) working in a model suggested by Willis (1922) managed to prove that assuming a single ancestral specie and probabilities of mutation and duplication a power law behavior is obtained. Expansion-modification systems proposed by Li (1991), which take into account the basic features of DNA mutation processes (Mansilla & Cocho, 2000), are also able to predict this behavior.

When discussing journal impact factors, a balance between the importance to the researchers of publish their work in high ranked journal, the difficulties associated with doing this and the increase of impact received by journals with high impact factor, seems to create a "rich gets richer" (the "Matthew Effect", see Merton, 1968; Egghe & Rousseau, 1990) mechanism also observed in the dynamics of complex networks (Barabasi, 2002). More than 49 years ago, Simon (1955, 1957) proposed a model which produces similar distributions. It is also interesting to notice that the bending of the tail of JIFs rank-order distribution means that after a critical zone of JIFs values is smooth thus discarding the possibility of the existence of multifractality.

Power-laws seem to be ubiquitous in physics, biology, geography, economics, linguistics, etc. (see Li, 2003). We consider "linguistic studies" not only those related with natural languages but also arbitrary languages over abstract finite alphabets. When the number of possible "words" is large, as it is the case for natural languages, it is expected to have a good fitness with a one-parameter power law. However, when the number of words is rather small, as it is the case of programming languages, one-exponent power laws absolutely fails and more parameters are necessary for a suitable fit. New elements to this considerations have been given by Le Quan et al. (2002). They showed that there is a serious deviation when the size of the sample is huge.

We expect that the increase in computing power will show that the deviation of Zipf's and Lotka's laws is a generic phenomenon. Then, a two-exponent Lotkaian and Zipfian Informetrics and linguistics should be welcome.

Acknowledgements

This work has been partially supported by the UNAM-PAPIIT grant IN-111003. The authors thank the sound comments of two anonymous referees.

References

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of $1/f$ noise. *Physical Review Letters,*, *59*, 381–384.
Barabasi, A. L. (2002). *Linked: The new science of networks*. Cambridge/Massachusetts: Perseus.
Condon, E. V. (1928). Statistics of vocabulary. *Science*, *67*, 300.

de Marchi, M., & Rocchi, M. (2001). The editorial policies of scientific journals: Testing an impact factor model. *Scientometrics*, *51*(2), 395–404.
Dewey, G. (1923). *Relative frequency of English speech sounds*. Cambridge/Massachusetts: Harvard University Press.
Dresden, A. (1922). A report on the scientific work of the Chicago section, 1897–1922. *Bulletin of the American Mathematical Society*, *28*, 303.
Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Amsterdam: Elsevier.
Egghe, L., & Rousseau, R. (1990). *Introduction de Informetrics*. Amsterdam: Elsevier.
Estoup, J. B. (1916). *Gammes stenographiques*. Paris: Institut Stenographique de France.
Fröhlich, G. (1996). The surplus value of scientific communication. *Review of Information Science*, *1*(2), 1–13.
Garfield, E. (1994). The impact factor. *Current Contents*, *25*, 3–7.
Langton, Ch. G. (1990). Computation at the edge of chaos. *Physica D*, *42*, 12–37.
Le Quan, H., Sicilia-García, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the 17th International Conference on Computer Linguistics*.
Li, W. (1991). Expansion-modification systems: A model for spatial $1/f$ spectra. *Physical Review E*, *43*, 5240.
Li, W. (2003). http://linkage.rockefeller.edu/wli/zipf/.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, *16*, 317.
Mandelbrot, B. B. (1954). Structure formelle des textes et communication. *Word*, *10*, 1–27.
Mansilla, R., & Cocho, G. (2000). Multiscaling in expansion-modification systems: An explanation for the long-range correlation in DNA. *Complex Systems*, *12*, 207.
McElvey, B. (2001). What is complexity science? *Emergence*, *3*, 137–157.
Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*, 56–63.
Nature. (2005). Editorial. *Nature, 435*, 1003–1004.
Popescu, I. (2003). On a Zipfs Law Extension to Impact Factors. *Glottometrics*, *6*, 83–93.
Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425.
Simon, H. (1957). *Models of man*. New York: Wiley and Sons.
Willis, J. (1922). *Age and area*. Cambridge, UK: Cambridge University Press.
Yule, G. (1924). A mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, *213*, 21–87.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge/Massachusetts: Addison-Wesley Press.