A geometric model for the analysis of citation distributions

Lucio Bertoli-Barsotti, Tommaso Lando

***Abstract*—** In this paper, we present an empirical study for modeling the citation distribution of papers of individual authors. We analyzed the citation records of applicants to the so called “Abilitazione Scientifica Nazionale” (ASN), a new procedure, based on scientific qualification criteria, for the recruitment of academic staff in Italy. We analyzed citation records of 131 physicists who were applicants in the ASN for a full professorship in the specific area of Condensed Matter Physics, using different mathematical models, namely: zeta, geometric, logarithmic and Pareto (of the first kind). Each model was “estimated”, on the basis of the observed citation pattern, via minimum Kullback-Leibler distance method. The geometric distribution was also considered by using a trimmed version of the estimator. As a measure of the effectiveness of the model, we computed the Kolmogorov-Smirnov distance. The most remarkable result is that the geometric distribution can provide an adequate tool for the modelization of the citation distribution of an author. Model fit may be further improved by adopting the trimming method.

***Keywords*—**Citation analysis, geometric distribution, trimmed estimator, Kullback-Leibler estimation, size-frequency data.

I. Introduction

Nowadays, research assessment is often based on citation counting techniques, used for evaluating scientific activities of individual researchers or research institutions. Number of citations received by articles, or individuals, are frequently used as a measure of “quality” in science. In citation analysis, there are two possible ways to interpret citation distributions and, accordingly, there are also two possible ways to fit citation distributions with probabilistic models, namely, i) the size-frequency or ii) the rank-frequency approach. In the first case, we consider the number of citations (for a given author) of each paper as observations that constitute a sample. Hence, the frequency of an observation c represents the number of articles with exactly c citations. Following a second approach, we may observe the rank of a paper and the frequency of its citations. In this case, the frequency of an observation, say r, is the number of citations of the paper ranked at the r-th position and we speak of rank-frequency distribution. In other words, we could interpret the citations as empirical observations (first approach) or frequencies (second approach). In other words, we could interpret the citations of papers as empirical observations (size-frequency case) or frequencies (rank-frequency case). In this paper we consider the problem form the point of view of the first approach, that is, the size-frequency analysis, therefore we try to answer to the question “Which is the best model for representing the citation-frequency curve?”

Several mathematical models have been proposed in the literature for the analysis of the citation distributions of the papers of different authors. Examples of models for citation distribution analysis are, for example, the Price distribution [1], exponential, stretched exponential (or Weibull distribution; [2]), Yule, Tsallis distribution, also known as q-exponential [3], [4], log-normal, discrete generalized beta distribution [5], [6], [7], [8], logarithmic distribution, to cite a few; and last but not least, a class of “Paretian” citation models which are of special interest in bibliometrics and in citation analysis.

The paper is organized as follows. In section II we present four different well known models and propose a general procedure to estimate their (unknown) parameters, in order to fit citation data as accurately as possible. Then, we present our dataset, which consists of 131 physicists who were applicants for a full professorship (in the specific area of Condensed Matter Physic) in Italy. Finally, we describe the results yielded by the estimation procedure in terms of goodness-of-fit between theoretical (estimated) and real (observed) data. The obtained results surprisingly show that, generally, the citation distributions of the considered authors comply to the geometric model, which can thereby represent a simple and valid alternative to other models, which are typically more popular in the field of bibliometrics.

II. Model distributions

We assume that, for a fixed author, the number of papers with c citations received is given by a formula $n(c,\boldsymbol{\alpha})$, $c = 1,2,\ldots$, where alpha is a (possibly vectorial) parameter. The unknown parameter alpha can be determined via a fitting procedure. By definition, the function n must satisfy the constraint $\sum_{c=1}^{\infty} n(c,\boldsymbol{\alpha}) = N$, where N represents the author’s total number of publications with at least one citation. We

This work has supported by the Italian funds ex MURST 60% 2014, 2015 and the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016 (to T.L.), supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

L.B.B. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; e-mail: lucio.bertoli-barsotti@unibg.it.

T.L. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolskà trida 33, Ostrava, Czech republic; e-mail: tommaso.lando@unibg.it.

might also give a probabilistic interpretation of n by noting that this function must be proportional to a probability mass function. Indeed

$$\sum_{c=1}^{\infty} f(c,\boldsymbol{\alpha}) = 1$$

where $f(c,\boldsymbol{\alpha}) = N^{-1}\cdot n(c,\boldsymbol{\alpha})$. In this form, f represents the probability mass function (p.m.f.) of a discrete random variable with support $\{1,2,3,\dots\}$.

In this paper we shall consider the following models of p.m.f..

A. *Zeta distribution*

The p.m.f. is $f(c,\alpha) = \frac{c^{-\alpha}}{\zeta(\alpha)}$, $\alpha > 1$, where $\zeta(\alpha) = \sum_{c=1}^{\infty} c^{-\alpha}$ denotes the Riemann zeta function ([9], p.527). This distribution is also referred to as *discrete Pareto distribution* and, depending on the context, it is also called *Zipf distribution* (see e.g. [10]). In the bibliometric literature, this formula is also known to as *power-law distribution*. When α is set equal to 2, $\zeta(2) = \frac{\pi^2}{6}$, we obtain the Lotka distribution [11], [12]. Somewhat strangely, in the literature the term Lotka's law is frequently used, in a more general sense than that used by Lotka [11], to refer to the above formula $c^{-\alpha}\zeta(\alpha)^{-1}$ as a size-frequency density function expressing the number/proportion of articles with exactly c citations [13], [14], [15], [16], [17], [18]. In applications, the "Lotka's law" is probably the simplest and the most used model for the analysis of citation frequency data.

B. *Geometric distribution*

We consider the p.m.f. of the geometric distribution in the following form: $f(c,\alpha) = \alpha(1-\alpha)^c$, $0 < \alpha < 1$. This model is also known as *shifted* geometric, because its support does not contains the value $c = 0$. The geometric distribution is an instance of a *power series distribution* (PSD, see [9]). A PSD follows the probability mass function of the type $b^{-1}\, a_c\theta^c$, for $c = 0,1,2,\dots$ where $a_c \geq 0$, and θ $(\theta > 0)$ is the so-called *power parameter*, and $b = \sum_{i=0}^{\infty} a_i\theta^i$ is the *series function*.

C. *Logarithmic distribution*

The p.m.f. is

$$f(c,\alpha) = -[\log(1-\alpha)]^{-1}c^{-1}\alpha^c,\ 0 < \alpha < 1$$

(see [9], p.302).

D. *Pareto distribution*

Let $f(c,\alpha) = \int_{c-0.5}^{c+0.5} l(y,\alpha)\,dy$, where

$$l(y,\alpha) = (\alpha-1)0.5^{\alpha-1}y^{-\alpha},\ y \geq 0.5,\ \alpha > 1$$

(see [19] p.574). The Pareto distribution is a continuous variant of the zeta distribution described above. For this reason, l is also known as continuous Lotka function (usually considered on the domain $[1,\infty)$). The model l is also known as the *Pareto distribution of the first kind* $P(I)(0.5,\alpha)$, where $\alpha > 1$ is the shape parameter [20]. Then, in this case the p.m.f. is obtained from the (continuous) distributional model $P(I)$, by continuity correction. In order to warrant the existence of its expectation, $\mu = \frac{\alpha-1}{\alpha-2}$, the condition $\alpha > 2$ must be assumed (unless to consider a truncated -i.e. with "cutoff"- version of the distribution).

Denote by c_i the number of citations gained by the i-th paper, $i = 1,\dots,N$. Let $C = \sum_{i=1}^{N} c_i$ the total number of citations (of an author). And let n_j the number of papers with exactly j citations. Let $F_N^*(t)$ be the empirical distribution function, defined as $F_N^*(t) = \sum_{j\leq t}\frac{n_j}{N}$, for every $t \in \mathbb{R}$.

Since in our context it is hard to assume the independence between observations, we rely on the estimation approach given by the minimum distance (MD) method (see [21], p.65-67; [22], [23]), by adopting the Kullback-Leibler distance. Remember that the mimimum distance estimate of the parameter $\boldsymbol{\alpha}$, with respect to the Kullback-Leibler distance, is the value of $\boldsymbol{\alpha}$ for which

$$-\sum_{j=1}^{c_{max}} n_j\ logf(j,\boldsymbol{\alpha}) = min_{\boldsymbol{\alpha}}\left(-\sum_{j=1}^{c_{max}} n_j\ logf(j,\boldsymbol{\alpha})\right).$$

It is worth noting that, under the independence assumption, the minimum Kullback-Leibler estimator coincides with the maximum likelihood estimator. Otherwise said, we search for the point $\widehat{\boldsymbol{\alpha}}$ for which the function $\sum_{j=1}^{c_{max}} n_j\ logf(j,\boldsymbol{\alpha})$ attains its absolute maximum value, given the set of observed pairs (j,n_j). All the models considered depend on two parameters, one for normalization (N), and one (α, as a scalar parameter) that characterizes the shape of the citation distribution.

III. Dataset

For a case study, we examined publication and citation data for applicants to the so called "Abilitazione Scientifica Nazionale" (ASN), a nation-wide evaluation based on scientific qualification criteria for the recruitment of academic staff in Italy. These data were also considered elsewhere for a comparative study concerning 13 different bibliometric indices [24],[25]. The ASN involved tens of thousands of candidates. Here we focus on its first edition, year 2012 (for candidates, the deadline for applications was November 20, 2012), so called ASN 2012.

The evaluation relied completely on applicants' research productivity (and it does not require any personal interaction between evaluators and candidates). An expert panel of evaluators (a Committee of five members) was asked to approve (''habilitate'') or to reject each candidate. In Italy, habilitation is necessary to be eligible for a full professorship.

Precisely, we consider a cohort of 131 physicists who were applicants in the ASN 2012 for a full professorship (from the original sample of 149 applicants, 18 scientists were discarded from the analyses due to insufficient citation data -e.g. an h-index less than 2- or difficulties in identifying the single scientist). The whole sample can be considered as highly homogeneous, in that information regarding individual publications are collected from a single well-defined area within Physics, i.e. Condensed Matter Physics, and all candidates must be considered on a similar level of scientific maturity and similar academic qualifications.

The publication and citation data were retrieved from Scopus in January 2014. Table I summarizes some of the most important citation metrics concerning our dataset: *C*=total number of citations; *N*=total number of cited papers; *C/N*=average number of citations per paper; c_{max} = citation count of the most cited paper; h=h-index g = g-index [26]; *R*= *R*–index [27]; basic statistics: *Mean* = Arithmetic mean; Q_i = *i*-th quartile; *Min* = minimum: *Max* = maximum; *SD* = standard deviation.

TABLE I: Main characteristics for the 131 datasets analyzed in the present study

	C	N	C/N	c_{max}	h	g	R
Mean	2206	85	25	359	21.6	39.7	37
Min	18	5	3	5	2	3	3.31
Max	13916	328	102	3068	53	200	102.1
Q1	1156	57	16	104	18	29	26.1
Q2	1786	77	21	177	22	40	36.7
Q3	2740	107	31	330	27	48.8	44.1
SD	1935	51	16	543	8.7	18.2	17.3

The applicants published a total of T=13347 papers, N=11079 of which cited at least once. The total number of citations was C=288972. We do not removed self-citations. The average percentage of uncited paper was 17%. The publications (cited at least once) received an average of 25 citations each (median = 23). The average percentage of citations of the most cited publication was 16%. The 77% of the authors received at least 1000 citations, and approximately 44% of the authors had at least 100 publications cited at least once. The most prolific author published 405 papers.

It is also of interest to analyze the dataset according to the values of the famous Hirsch index h, defined as the maximum number of articles with at least h citations each, and other h-type indices, viz., the g- and R-indices. Generally, h-type indices are aimed at assessing scientists, by summarizing (based on citation data) both their productivity and impact on the scientific community by a single number. The applicants' h-index values were on average 21.6 (median = 22), and ranged from a minimum of 2 to a maximum of 53. We found an average h-index of 21.6. The maximum observed value for h was 53. In contrast, only 13% of the scientists have an h-index smaller than 10. As for the other considered h-type indices, Egghe [26] defines the g-index as follows: "The highest number g of papers that together received g^2 or more citations". Differently, the *R*–index is defined as the square root of the sum of citations in the Hirsch core (that is, the set consisting of the most cited h papers). The *R*–index and the g–index present very similar results, as already pointed out in [28], and are both highly correlated with c_{max}. Then, it may be argued that their values could be inflated by a single highly cited paper. Besides, it is known that the h-index is positively correlated with the total number of citations *C* as well as to the number of publications *N* [29]. On the other hand, increasing publications *alone*, or the total number of citations *alone*, or *C/N alone*, does not have immediate effect on the h-index. Summarizing, the h-index combines in a simple way both productivity (*N*) and quality (*C/N*), but it is relatively insensitive to one (or few) very highly cited papers and, at the same time, it does not take into account the citation counts of papers with fewer than h citations. In short, the h-index captures only "a small amount of information about the distribution of a scientist's citations" ([30], p. 2).

IV. RESULTS

As discussed above, the main goal of this analysis was to obtain an estimate of the number of papers with c citations, for every $c \in \{1,2,\dots,c_{max}\}$, using a theoretical model distribution. Based on the empirical observations, we estimate each one of the considered models and we compute a Kolmogorov-Smirnov (K-S) distance as a discrepancy measure (between observed and fitted data) for goodness-of-fit purposes.

Indeed, the K-S distance can be used (here only for descriptive and comparative purposes) to compare the different distributional assumptions and to identify the model which better complies with observed citation frequency data, among those considered. We recall that the K-S statistic, say D_N, is defined as the maximum (vertical) distance between the empirical and the estimated theoretical distribution function $F_{\hat{\alpha}}(t)$, that is, in symbols, $D_N = \sup_t |F_{\hat{\alpha}}(t) - F_N^*(t)|$. For taking into account the sample dimension, we also compute the statistic $D_N^* = D\sqrt{N}$, more useful for comparisons between different sizes.

We observe that, frequently, the sets of citations contain outliers, that is, some author may have one (or few) article(s) which has been cited an outstandingly high number of times, compared to all his (or her) other papers. This may be due to several reasons ("age" of the paper, number of co-authors, etc.). We note that the presence of outliers has a negative influence on the geometric model. Table 2 reports the correlation coefficients between the K-S distance and some of the most important bibliometric indicators, namely: the h-index; the total number of citations *C*; the number of papers with at least one citation, *N*; the average number of citations per paper *C/N* and the number of citations of the most cited paper c_{max}. From data reported in Table 2, we can observe a quite strong dependence between the K-S distance between empirical and geometric distribution and the maximum number of citations (c_{max}), and thereby the average number of papers (*C/N*). This result suggest that, for the geometric model, the goodness-of-fit could be enhanced by excluding from the sample the highest observed values, which can actually be considered as outliers. In particular, we find a satisfactory improvement of the goodness of fit by trimming the 5% of the highest observations in each sample dataset (note that, however, the K-S distance is evaluated over the whole sample). As the MD estimator for the geometric distribution is the reciprocal of the sample mean, then we estimate α by the reciprocal of a trimmed (truncated) sample mean, which is less sensitive to outliers and is especially suitable for dealing with heavy tailed distributions. However, we observed that this technique has a negative effect on the other models, as also confirmed by Table II, where: *zeta*= zeta distribution; *geo*=geometric distribution; *log*=logarithmic distribution; *par*=Pareto distribution; *geo(t)*= geometric distribution with a trimmed estimate of its parameter.

Interestingly, the logarithmic model yields even a better fit for higher values of c_{max}. For these reasons, we report the values produced by the trimming method, just for the geometric distribution (*geo(t)*).

TABLE II: Correlations between the K-S distance and some important citation metrics

	zeta	*geo*	*log*	*par*	*geo(t)*
h	0.17	0.10	0.02	-0.06	-0.36
C	0.14	0.31	-0.07	-0.08	-0.29
n	0	-0.07	-0.13	-0.22	-0.46
C/n	0.28	0.50	0.08	0.06	0.05
MC	0.16	0.60	-0.11	-0.05	-0.03

From Table II we observe that the geometric distribution, estimated with the trimming method, is substantially insensitive to the citation count of the most cited paper c_{max} and *C/N*, but is positively influenced by the bibliometric indicators of “productivity” (h, *C* and *N*). In particular, the K-S distance is reduced for larger samples (a sort of consistency), especially with the (trimmed) geometric distribution.

Table III reports, in the first row, the average value $M(D)$ of the K-S statistic D, over the 131 datasets, and, in the second, third and fourth row, respectively, the number of cases in which this average is smaller than 0.1, 0.15 and 0.2. For completeness, we also considered the finite size version of the Zipf distribution with a finite range $\{1,2,\dots,c_{max}\}$. Introducing the upper limit (cut-off) c_{max} for the summation, we obtain the so called *Estoup* distribution (*est* in the Table III), with the citation distribution function $N\sum_{c=1}^{c_{max}}\frac{c^{-\alpha}}{\zeta(\alpha)}$.

As can be seen, the geometric model shows better fit compared to the other distributions, especially when the parameter is estimated with the trimmed MD estimator. Moreover, note that using a finite size version of the Zipf model gives an improvement to the quality of the fit. Indeed, we observe that the Estoup distribution is well fitting to data, compared to the non-truncated zeta distribution. We recall that the zeta distribution is defined only when the parameter α is greater than 1, while the Estoup distribution allows the parameter α to be less or equal to 1. In particular, in our estimation results we observe an average value of $\hat{\alpha}$ of 1.37 for the zeta model and 1 for the Estoup model, for which, for 51 over 131 datasets, $\hat{\alpha}$ is less than 1. It should be stressed, that the condition $\alpha < 2$, for the zeta distribution, implies the non-existence of the first moment, which means, in this case, that the expected average number of citations of an author is infinite, as well as the expected total number of citations. Therefore, we argue that citation-frequency profiles may comply to a power law, provided that the truncated version (i.e. the Estoup distribution) is considered.

TABLE III: Values of the K-S distance for different model distributions

	zeta	*geo*	*log*	*est*	*par*	*geo(t)*
M(*D*)	0.26	0.18	0.18	0.14	0.25	0.12
#(*D*<0.10)	1 (1%)	19 (14%)	10 (7%)	23 (17%)	0 (0%)	52 (40%)
#(*D*<0.15)	2 (1%)	49 (37%)	44 (33%)	77 (58%)	0 (0%)	105 (80%)
#(*D*<0.20)	12 (9%)	83 (63%)	80 (61%)	121 (93%)	56 (43%)	123 (94%)

Finally, Table IV summarizes the basic statistics regarding the metric D^* for the whole sample of datasets. The geometric distribution estimated via trimming method resulted the most reliable among the considered models. This is also shown in Fig. 1, which compares the fitted distributions yielded by the geometric and zeta models with the empirical distribution of a Physicist, who is particularly suitable for representing the whole sample, in terms of total number of citations (2091), cited papers (89) (which are definitely on average) and h-index (26). As apparent from Fig. 1, the geometric model is extremely well-fitting to the empirical distribution, compared to the zeta model.

TABLE IV: Summary statistics concerning the metric $D_N^* = D\sqrt{N}$

*D**	*zeta*	*geo*	*log*	*est*	*par*	*geo(t)*
Mean	2.26	1.60	1.58	1.24	1.81	1.00
Min	0.29	0.26	0.17	0.25	0.53	0.32
Max	4.59	4.89	3.34	2.98	3.31	2.49
SD	0.79	0.84	0.65	0.53	0.51	0.35
Q1	1.76	1.02	1.09	0.81	1.50	0.77
Q2	2.30	1.47	1.58	1.24	1.81	0.95
Q3	2.74	2.06	2.02	1.54	2.10	1.18

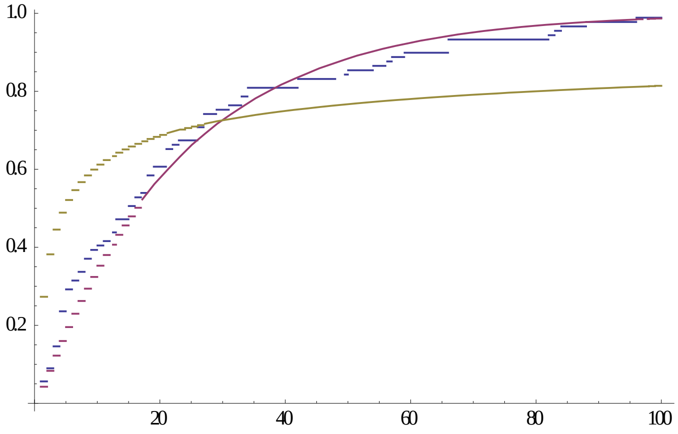

Fig. 1. Empirical and estimated distributions with the geometric and zeta models

V. Conclusion

In this paper we considered four different types of distributions, suitable for describing citation frequency data -

that is: zeta; geometric; logarithmic and Pareto (of the first kind). All these models–that are directly comparable because are families that depend on a single scalar parameter (a shape parameter α)–are used for fitting the citation frequency curves of a relatively homogeneous cohort of physicists. The investigated scientists can be considered as "average authors" (the mean value of the *h*-index was about 21), then our case study can be considered less typical than would be expected from a standard informetric analysis, where attention is focused on very prominent persons.

Precisely, we analyzed the citation records of 131 physicists who were applicants in the ASN 2012 for a full professorship. Each one of the models considered was estimated, on the basis of the observed citation pattern, via MD (Kullback-Leibler) method. The geometric distribution was also considered by using a trimmed version of the MD estimator.

Overall, our study provides sufficient evidence of a somewhat unexpected result: the (perhaps) most popular model for the analysis of citation frequency data -i.e. the Pareto distribution of the first kind, known as the Lotka's power law- is not always the best candidate for the representation of the citation frequency curve.

Analysis of the data of citation distribution of papers of different authors revealed that, at least as far as concerns the size-frequency data at hand, the geometric distribution was found to be a better alternative to more traditional models (for evaluating the quality of the fit, we used the K-S statistic). It is interesting to note that the geometric model represents the data more satisfactorily when the trimmed version of the Kullback-Leibler estimator is adopted. Further studies will be conducted to verify our results.

REFERENCES

[1] W. Glänzel, "On the H-index – A mathematical approach to a new measure of publication activity and citation impact," *Scientometrics* Vol. 67, No. 2, pp. 315-321, 2006.

[2] J. Laherrere and D. Sornette, "Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales," *European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 2, No. 4, pp. 525–539, 1998.

[3] C. Tsallis, and M.P.de Albuquerque, "Are citations of scientific papers a case of nonextensivity?," *European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 13, No. 4, pp. 777–780, 2000.

[4] Q.L. Burrell, "Extending Lotkaian Informetrics.," *Information Processing and Management*, Vol. 44, No. 5, pp.1794–1807, 2008.

[5] G. Martinez-Mekler, R.A.,Martinez, M.B. del Rio,. R.,Mansilla, P. Miramontes, and G., Cocho, "Universality of rank-ordering distributions in the arts and sciences," *PLoS ONE*, Vol. 4, e4791, 2009.

[6] A.M. Petersen H.E., Stanley, and S., Succi, "Statistical regularities in the rank-citation profile of scientists," *Sci Rep*, 2011

[7] J.M. Campanario, "Distribution of ranks of articles and citations in journals," *Journal of the American association for information science and technology*, Vol. 61, No. 2, pp. 419–423, 2010.

[8] R. Mansilla E, Köppen G, Cocho, and P. Miramontes, "On the behavior of journal impact factor rank-order distribution," *Journal of Informetrics*, Vol. 1, pp. 155–160, 2007.

[9] N.L. Johnson, A.W. Kemp and S. Kotz., *Univariate discrete distributions*, Wiley Series in Probability and Statistics (third ed.) John Wiley, New York, 2005.

[10] P. T. Nicholls, "Empirical validation of Lotka's law," *Information Processing and Management*, 22, pp. 417-419, 1986.

[11] A. J. Lotka, "The frequency distribution of scientific productivity," *Journal of the Washington Academy of Sciences*, vol. 16, no. 12, pp. 317-323, 1926.

[12] R. C. Coile, "Lotka's frequency distribution of scientific productivity," *Journal of the American Society for Information Science*, vol.28, no.6, pp.366-370, 1977.

[13] L. Egghe, *Power laws in the information production process: Lotkaian informetrics*, London: Academic Press, 2005.

[14] L. Egghe, "Relations between the continuous and the discrete Lotka power function," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 7, pp. 664–668, 2005.

[15] L. Egghe, R. Rousseau, "An informetric model for the Hirsch-index", Scientometrics, vol. 69, no. 1, pp. 121–129, 2006.

[16] L. Egghe, "Lotkaian informetrics and applications to social networks", *Bulletin of the Belgian Mathematical Society-Simon Stevin*, vol. 16, no. 4, pp. 689–703, 2009.

[17] L. Egghe, "A new short proof of Naranan's theorem, explaining Lotka's law and Zipf's law," *Journal of the American Society for Information*, vol. 61, no. 12, pp. 2581-2583, 2010.

[18] B. Rousseau, R. Rousseau, " LOTKA: A program to fit a power law distribution to observed frequency data," Cybermetrics, vol. 4, no.1.

[19] N. L. Johnson, S. Kotz and N. Balakrishnan, *Continuous univariate distributions*, Vol. 1, 2nd Edition. John Wiley, New York, 1994

[20] B. C. Arnold, *Pareto Distributions*, International Cooperative Publishing House, Fairland, Maryland, 1983.

[21] A. A. Borovkov, *Mathematical Statistics*, Amsterdam: Gordon and Breach Science Publishers, 1998.

[22] T. Lando and L. Bertoli-Barsotti, "Statistical Functionals Consistent with a Weak Relative Majorization Ordering: Applications to the Mimimum Divergence Estimation," *WSEAS Transactions on Mathematics*, Vol. 13, p. 666-675, 2014.

[23] Lando, T., Bertoli-Barsotti, L., 2014, Divergence Measures and Weak Majorization in Estimation Problems, in *Advances in Applied and Pure Mathematics - 2nd International Conference on Mathematical, Computational and Statistical Sciences (MCSS '14)*, May 15-17, 2014, Gdansk, Poland, p.152-157.

[24] T. Lando, L. Bertoli-Barsotti, "A New Bibliometric Index Based on the Shape of the Citation Distribution," PLoS ONE, vol. 9, no. 12: e115962, 2014.

[25] Bertoli-Barsotti, L., Lando, T., 2015, Informetric models for citation frequency data: an empirical investigation, in Nikos E. Mastorakis, Panos M. Pardalo and Ravi P. Agarwal (Eds.), *New Developments in Pure and Applied Mathematics - Proceedings of the International Conference on Mathematical Methods, Mathematical Models and Simulation in Science and Engineering (MMSSE 2015)*, Vienna, Austria, March 15-17, 2015, p. 37-39.

[26] L. Egghe, "An improvement of the *h*-index: The g-index," *ISSI Newsletter*, Vol. 2, No. 1, pp. 8–9, 2006.

[27] B.H. Jin, LM. Liang, R. Rousseau, and L. Egghe, "The R- and AR-indices: complementing the h-index," *Chinese Science Bulletin*, Vol. 52, No. 6, pp.885–863, 2007..

[28] A. De Visscher, "What Does the g-Index Really Measure? ," *Journal of the Association for Information Science and Technology*, Vol. 62, No. 11, pp.2290–2293, 2011.

[29] van Raan, A.F.J. (2006). Statistical Properties of Bibliometric Indicators: Research Group Indicator Distributions and Correlations. *Journal of the American Society for Information Science and Technology* 57, 3, 408-430.

[30] Joint Committee on Quantitative Assessment of Research. (2008). Citation statistics. A report from the International Mathematical Union (IMU) in cooperationwith the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). Berlin, Germany: International Mathematical Union (IMU).