Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 185–192
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Scoring research output using statistical quantile plotting

Jan Beirlant [a], Wolfgang Glänzel [b], An Carbonez [c,*], Herlinde Leemans [d]

[a] *Department of Mathematics and University Center of Statistics, K.U. Leuven, Belgium*
[b] *Faculty of Economics and Steunpunt O&O Statistieken, K.U. Leuven, Belgium*
[c] *University Center of Statistics, K.U. Leuven, Belgium*
[d] *Dienst Onderzoekscoordinatie, K.U. Leuven, Belgium*

Received 21 November 2006; received in revised form 20 April 2007; accepted 23 April 2007

Abstract

In this paper, we propose two methods for scoring scientific output based on statistical quantile plotting. First, a rescaling of journal impact factors for scoring scientific output on a macro level is proposed. It is based on normal quantile plotting which allows to transform impact data over several subject categories to a standardized distribution. This can be used in comparing scientific output of larger entities such as departments working in quite different areas of research. Next, as an alternative to the Hirsch index [Hirsch, J.E. (2005). An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572], the extreme value index is proposed as an indicator for assessment of the research performance of individual scientists. In case of Lotkaian–Zipf–Pareto behaviour of citation counts of an individual, the extreme value index can be interpreted as the slope in a Pareto–Zipf quantile plot. This index, in contrast to the Hirsch index, is not influenced by the number of publications but stresses the decay of the statistical tail of citation counts. It appears to be much less sensitive to the science field than the Hirsch index.

Keywords: Quantile plots; Standardizing; Normal quantile plot; Pareto quantile plot; Extreme value index

1. Introduction

Rating and comparing research groups on a macro level based on some kind of quality indicator concerning the journals in which they publish is regularly used in university policy decision-making. Such kinds of methods are often criticized, especially because the indicators are strongly dependent on the field of research. This is especially the case when using impact factors. Limitations of impact factors are, for instance, discussed in Moed (2002), Glänzel and Moed (2002) and Podlubny (2005). A simple rating measure of research groups such as on departmental level, could be the count of papers in those journals which are situated within the top 10% of a (S)SCI subject category. Such a scoring method could be a part of a management action to stimulate publication in top journals rather than stimulating a higher number of publications. This top 10% scoring is in many ways unsatisfactory due to the volatility of a journal impact factor over the years together with the problems linked to the composition of a subject category. Here we present a scoring technique, to be used only in macro level comparisons, characterized by standardization between different

* Corresponding author. Tel.: +32 16 32 22 42; fax: +32 16 32 28 31.
E-mail address: An.Carbonez@ucs.kuleuven.be (A. Carbonez).

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.04.002

subject categories, and where possible negative effects due to the composition of a subject category and the volatility of impact factors are somehow relaxed due to the continuity in the score (in contrast to a top 10% rule where 0–1 scores are attributed to journals). Of course this method would also benefit from its use on homogeneous composition of subject categories on which it is based.

Next, we also present a scoring method for individual researchers. Here recently the Hirsch index constitutes a recent proposal (Hirsch, 2005). However, it turns out to be quite dependent on the field of research and the number of papers a researcher has published, rather than scoring the intrinsic interest of the research community in the contents of the papers. In case of Lotkaian–Zipf–Pareto behaviour of citation counts of an individual, the Pareto tail index can be interpreted as the slope in a Pareto–Zipf quantile plot. More generally, one could use the extreme value index of the statistical distribution of the citations of an individual. This concept generalizes the Pareto tail index in cases of non-Lotkaian (or non-Zipf/Pareto) behaviour.

We start by considering a set of impact factors from a list of journals as a realization of a random variable. The same can be done for the number of citations of the different papers of an individual at a given moment. Both solutions are based on the concept of quantile plots. The use of quantile plots in informetrics investigations is illustrated in Huber (1998). In a quantile–quantile plot (QQ plot), the quantiles of the data are plotted against the quantiles of a specific distribution (assuming that the data follow the specific distribution). The basic idea is that, if the data follow that specific distribution, then the graph will essentially be a straight line. See Gilchrist (2000) for a comprehensive reference on QQ plots. Many statistical packages (including Excel) have quantile functions available for many standard distributions.

The statistical distribution of impact factors of a set of journals can be well approximated by a lognormal distribution. This assertion can be verified using a normal QQ plot or a goodness-of-fit test. Let us remind that a quantile function taken at a value $p \in (0,1)$ yields the $100p$-percentile of the corresponding distribution. The ith smallest observation $X_{i,n}$ $(1 \leq i \leq n)$ from a sample of size n can be considered as the $100(i/n)$-percentile of the data distribution. Given a sample of size n (say impact factors of n scientific journals), the normal QQ plot is then defined by

$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right), X_{i,n}\right), \quad i = 1, \ldots, n \tag{1}$$

where Φ^{-1} denotes the standard normal quantile function. Here one typically applies a continuity correction $i/(n+1)$ to the fraction i/n. Another possible choice is $(i-0.5)/n$. Because any normal random variable X with mean μ and variance σ^2 possesses the same distribution as $\mu + \sigma Z$ where Z is standard normally distributed, one has that a normal QQ plot is approximately linear in case the data constitute a normal random sample. More importantly, this plot defines a transformation to standard normality. Indeed if the normal QQ plot for instance is of exponential type then a logarithmic transformation of the data leads to a linear QQ plot, so that then the original data are lognormally distributed.

This can approximately be observed in case of the *JCR Science Edition 2005* impact factors for the subject categories *Statistics and probability* (with $n = 81$ journals), *Biochemistry and molecular biology* (with $n = 356$ journals), and *Medicine, research and experimental* (with $n = 89$ journals). In Fig. 1 the normal QQ plots of the original impact factors and log-transformed impact factors of these two subject categories are presented, together with a histogram representation of the data. Remark the differences between these two distributions: the maximum impact factors, respectively, are 8.4, 74.4, and 40.2, while the averages, respectively, are 1.5, 6.2 and 4.5.

In Section 2, we will present a scoring method transforming the different distributions towards the same lognormal distribution.

Section 3 deals with the problem of scoring the publication output of individual researchers. Here the distributional family that could serve as a reference is the Lotka–Zipf–Pareto distribution. Here we can refer to Glänzel (2006) and Egghe and Rousseau (2006). The strict Pareto distribution is defined by

$$P(X > x) = \left(\frac{x}{t}\right)^{-\alpha}, \quad x \geq t, \tag{2}$$

with $\alpha > 0$ and $t > 0$. As is shown in Beirlant, Goegebeur, Segers and Teugels (2004), this distribution is characterized by the fact that a log-transformed Pareto random variable Y is exponentially distributed with survival function $P(Y > y) = \mathrm{e}^{-\alpha(y - \log t)}$ $(y > \log t)$ and quantile function $Q(p) = \log t - (1/\alpha)\log(1-p)$ $(p \in (0,1))$, such that a Pareto QQ

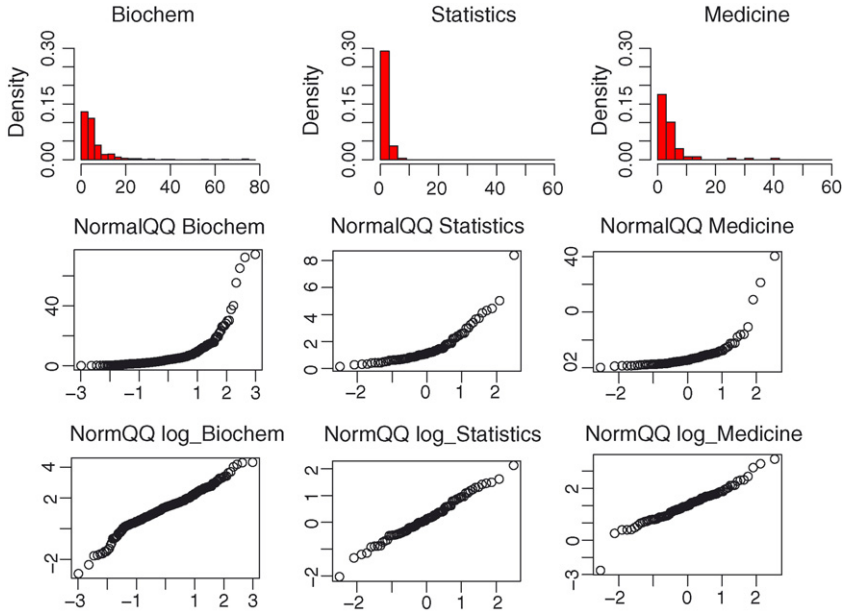

Fig. 1. Histograms, normal QQ plots of the impact factors, and normal QQ plots of the log-transformed impact factors for the subject category Biochemistry and Molecular Biology (left), Probability and Statistics (middle), and Medicine, Research and Experimental (right).

plot can be constructed as an exponential QQ plot based on the log-transformed data:

$$\left(-\log\left(1-\frac{i}{n+1}\right), \log X_{i,n}\right), \quad i=1,\ldots,n. \tag{3}$$

In case of linearity remark that the slope of a Pareto QQ plot approximates the Pareto tail index $1/\alpha$. Pareto behaviour can approximately be observed for instance with most of the Price Medal awardees from the international journal *Scientometrics*. In Fig. 2, the Pareto QQ plots are given for the citation counts for E. Garfield and another Price Medallist with the same Hirsch index as reported in Glänzel and Persson (2005). Both plots are based on their highest 17 citation counts. The Hill estimator for the slope discussed in Section 3 is close to 1 in case of E. Garfield, while the other plot induces a slope estimate close to 0.4. Practically speaking the value 1 should be regarded as a maximum since a value ≥ 1 corresponds to a distribution with an infinite mean, while a slope >0.5 corresponds to an infinite variance. The Pareto tail index, also termed the extreme value index, will be presented as an index with which to compare individual researchers.

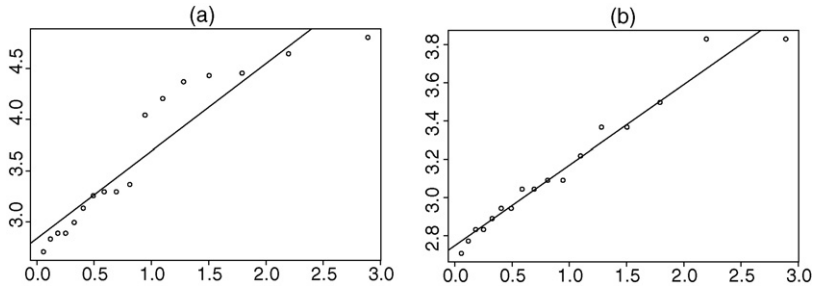

Fig. 2. Pareto QQ plots based on the highest 17 citation counts for (a) E. Garfield and (b) another Price Medallist with the same Hirsch index, with fitted least squares line.

2. Rescaled journal impact factors

In Section 1, it was mentioned that a normal QQ plot can be used to deduce a transformation towards normality. However, it can also be used to transform the data in a nonparametric way without using a specific mathematical function as a logarithmic function. Indeed, as it is used in the nonparametric van der Waerden test (see, for instance Conover, 1999) for comparing different data groups,

$$X_{i,n} \to \Phi^{-1}\left(\frac{i}{n+1}\right), \quad i = 1, \ldots, n \tag{4}$$

transforms an order statistic in a standard normal score. Or, when introducing the rank R_i of an observation X_i, any observation can be substituted by

$$X_i \to \Phi^{-1}\left(\frac{R_i}{n+1}\right), \quad i = 1, \ldots, n. \tag{5}$$

Hence whatever the underlying distribution of the data, the van der Waerden scores are standard normally distributed. Transforming these scores with an exponential function results in a score, the distribution of which is of the lognormal type as with the original impact factor distributions:

$$X_i \to a^{\Phi^{-1}}\left(\frac{R_i}{n+1}\right), \quad i = 1, \ldots, n \tag{6}$$

for some base number a.

The characteristics of the resulting distribution of course depend on the value a. In Fig. 3, the resulting scoring distribution is shown for the Statistics and Probability SCI subject category when using $a = 1.7$.

One could limit the distribution of the transformed scores to the right, for instance by attributing a score 1 to the top 10% journals. This then limits the impact of the highest impact journals. Also the distributions of the different transformed scores over the different subject categories only differ in the length of the tails of these distributions due to its dependence on n. This is most pronounced in the right hand side tail. This second-stage score is defined by

$$X_i \to a^{(\Phi^{-1}(R_i/(n+1)) - \Phi^{-1}(0.9))_-}, \quad i = 1, \ldots, n \tag{7}$$

where $u_- = u$ when $u < 0$ and $u_- = 0$ when $u > 0$.

The results for the scoring method (7) for the three considered subject SCI categories are also shown in Fig. 4 when using $a = 1.7$.

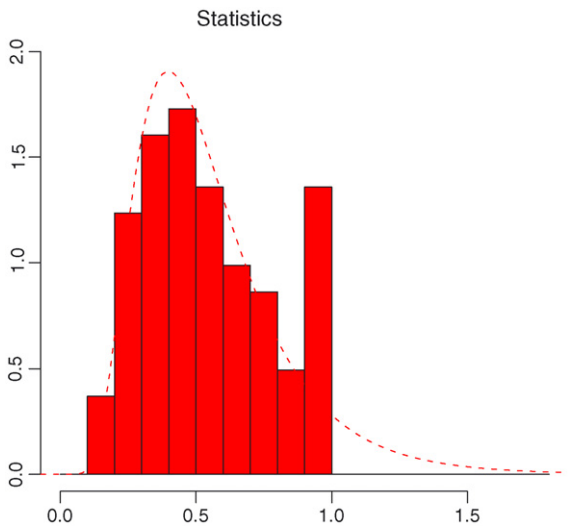

Fig. 3. Distribution of the transformed scores for the Statistics and Probability SCI subject category.

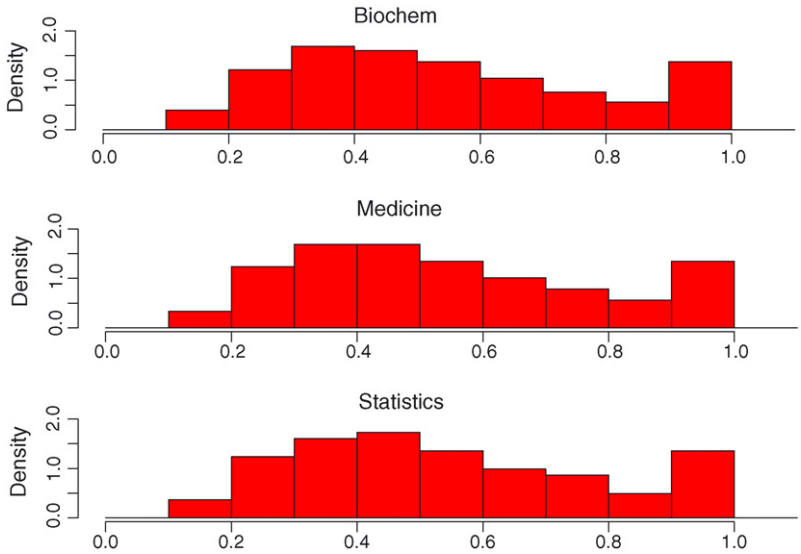

Fig. 4. Distributions of the transformed scores (7) for three SCI subject categories.

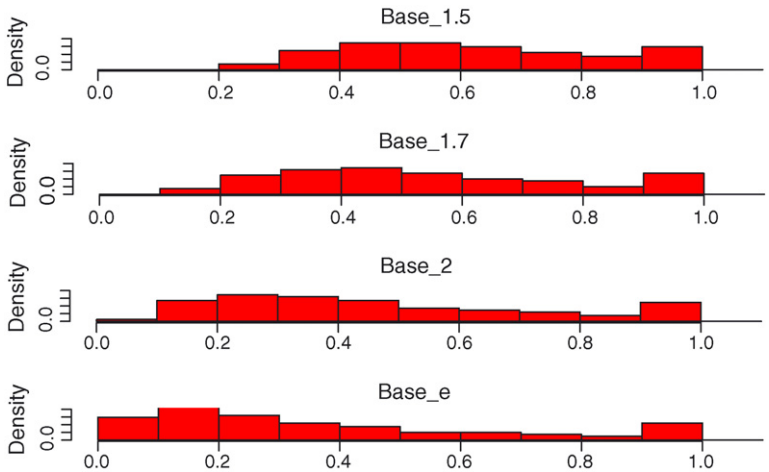

Fig. 5. Distributions of the transformed scores (7) for different values of a.

The choice of a can best be decided upon by the user. Table below can help in deciding upon this value. The mean, median, 75 and 90 percentiles of the resulting distribution are given in Table. In Fig. 5 these four distributions are represented with a histogram view.

a	Mean	Median	P75	P90
1.5	0.623	0.59	0.78	0.98
1.7	0.551	0.51	0.72	0.97
2	0.477	0.41	0.65	0.96
e	0.378	0.28	0.53	0.95

We end this section by an application to the analysis of the research output of all Belgian mathematical sciences researchers over the period 1996–2004. Indeed the scoring algorithm can not only be used to standardize the impact factors between different subject categories, but also to make comparisons possible for one subject over time. The transformed scores over this period for the mathematical sciences (comprising among others Mathematics, pure and

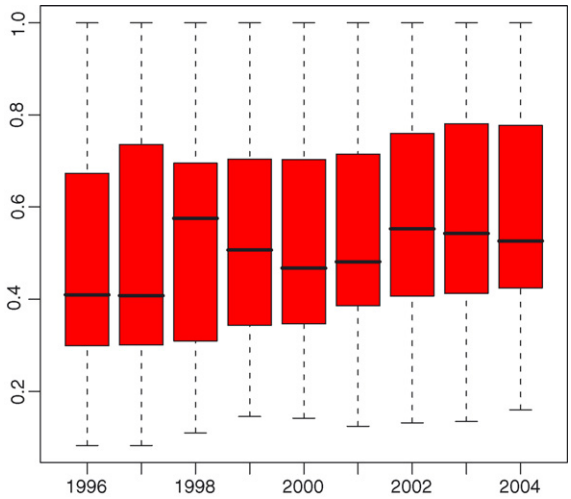

Fig. 6. Boxplots of the transformed scores (7) for the papers of the Belgian mathematical sciences researchers for the period 1996–2004.

applied, and Statistics and Probability) were applied to the SCI paper output in Belgium. In Fig. 6, the scores for the Belgian mathematicians are given using boxplots. It can be observed that the standardized scores have slightly risen over the given period.

3. Assessing individual research performance using the Pareto tail index

In the introductory section based on a simple Pareto model (2) the extreme value index $1/\alpha$ was proposed as an indicator for research performance of an individual researcher. However, model (2) can only be observed when restricting on some subset of the papers with the highest number of citations, say $X_{n-k,n} \leq X_{n-k+1,n} \leq \cdots \leq X_{n,n}$ for some $k \in \{1, \ldots, n\}$. In general however one has that

$$\lim_{x\to\infty} x^{\alpha} P(X > x) = C \tag{8}$$

for some constant $C > 0$. Under this model Glänzel (2006) has shown that the Hirsch index H asymptotically for n large satisfies $H = An^{1/(\alpha+1)}(1 + o(1))$ for some constant $A > 0$. In contrast the index $1/\alpha$ can be regarded as an index of research performance which makes abstraction of the number n of published papers.

In the literature on extreme value theory as for instance outlined in Embrechts Kliippelberg and Mikosch (1997) or Beirlant et al. (2004), it is shown that this tail-Pareto model can best be respecified as

$$\lim_{x\to\infty} \frac{P(X > tx)}{P(X > t)} = x^{-\alpha}, \quad x > 1, \tag{9}$$

or equivalently

$$P(X > x) = x^{-\alpha} l(x) \tag{10}$$

for some slowly varying function l satisfying $\lim_{t\to\infty} l(tx)/l(t) = 1$. A popular submodel of (8), (9) or (10) is given by

$$P(X > x) = Cx^{-\alpha}(1 + Dx^{-\beta}(1 + o(1))) \text{ as } x \to \infty \tag{11}$$

for some constants $C > 0$ and D, and with some $\beta > 0$. Samples from such models in a Pareto QQ plot only show an ultimate linear behaviour at the highest observations. This happens typically with distributions of citation counts as is illustrated in Fig. 7 showing the Pareto QQ plot with the citations for all papers of two Scientometrics Price Medallists.

The estimation of α under this Pareto-type model has received considerable attention. Here the problem of the selection of a subset $X_{n-k,n} \leq X_{n-k+1,n} \leq \cdots \leq X_{n,n}$ is a key problem in this field. The estimator of $1/\alpha$ under (9)–(10)

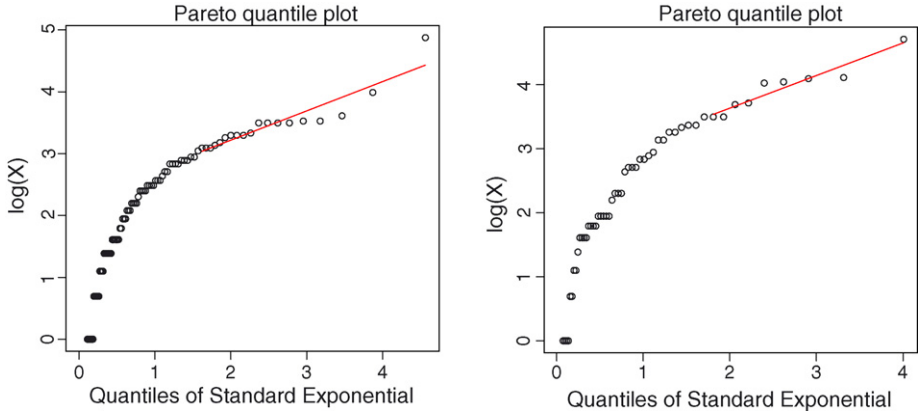

Fig. 7. Pareto quantile plots for the citation counts of all papers of two Price Medallists, with fitted regression lines at the values k indicated in Fig. 6.

which has received most attention was proposed by Hill (1975):

$$H_{k,n} = k^{-1}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n}. \tag{12}$$

It can be regarded as a maximum likelihood estimator, and as a naive estimator of the slope of the Pareto QQ plot to the right of the anchor point at the height $\log X_{n-k,n}$. It is also the mean excess value of the log-transformed data above $\log X_{n-k,n}$. $H_{k,n}$ is asymptotically normally distributed (when k, $n \to \infty$ and $k/n \to 0$) with variance $1/(k\alpha^2)$ which allows to construct confidence intervals for $1/\alpha$. For more details we refer to Chapter 4 in Beirlant et al. (2004). The values of $H_{k,n}$, which are called the EVI (Extreme Value Index), are plotted against k as shown in Fig. 8 for the QQ plots in Fig. 7.

In order to choose k adaptively several algorithms have been proposed in literature. Most algorithms aim at finding the value of k for which the asymptotic mean squared error (AMSE) of $H_{k,n}$ is minimal. The AMSE of the Hill estimator in case of the model (11) is given by

$$\mathrm{AMSE}(H_{k,n}) = \frac{1}{k\alpha^2} + \left(\frac{D\beta(k/n)^{\beta}}{1+\beta}\right)^2.$$

This expression can be estimated as shown in Beirlant, Dierckx, Guillou and Starica (2002), leading to a selection rule $\hat{k}$. The code for the selection algorithm in *S Plus* can be found on http://ucs.kuleuven.be/Wiley/index.html. These values of $\hat{k}$ can be found in Fig. 8 with a vertical line leading in both cases to values $H_{\hat{k},n}$ around 0.4. Other algorithms for selecting k, together with extensions of these methods to non-Pareto tails can be found in Chapters 4 and 6 in Beirlant et al. (2004).

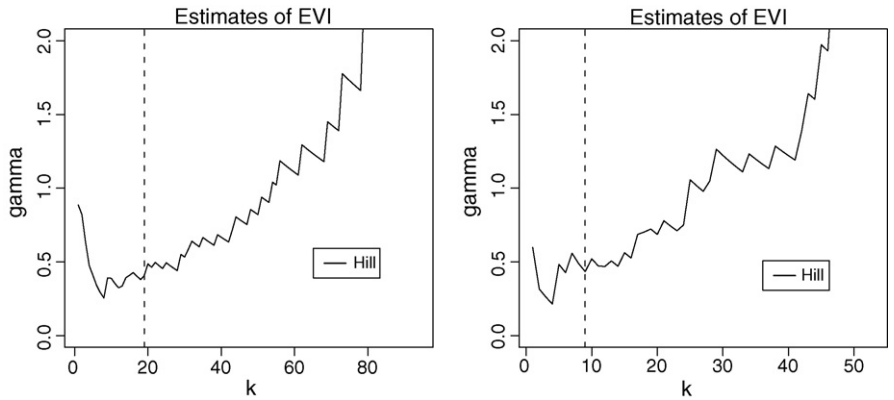

Fig. 8. Plots of the Hill estimates as a function of k corresponding to the citation counts of all papers of two Price Medallists.

4. Conclusion

In this paper, we have presented two scoring methods for assessing the scientific output, first on a macro and meso level for comparing different groups of researchers based on indicators such as impact factors, and secondly for comparing individual researchers based on the extreme value index. Both proposals are based on quantile plotting.

The second proposal can use some extra research. Questions such as the detection of non-Paretian behaviour in citation counts in an automatic way, next to the modelling of the index for a researcher over time, and the comparison of distributions of the Hill estimates for researchers from different research fields are of special interest. The ultimate top researchers, however, in any given field will exhibit an extreme value index close to 1.

References

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2004). *Statistics of extremes: Theory and application*. Wiley., pp. 490

Beirlant, J., Dierckx, G., Guillou, A., & Starica, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, *2*, 157–180.

Conover, W. J. (1999). *Practical nonparameteric statistics* (3rd ed.). New York: Wiley & Sons., pp. 584.

Embrechts, P., Kliippelberg, C., & Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Berlin: Springer Verlag.

Gilchrist, W. G. (2000). *Statistical modelling with quantile functions*. Chapman & Hall/CRC.

Glänzel, W. (2006). On the H-index a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.

Glänzel, W., & Persson, O. (2005). H-index for price medallists. *ISSI Newsletter*, *1*(4), 15–18.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann of Statistics*, *3*, 1163–1174.

Hirsch, J. E. (2005). An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572 (also available at: http://arXiv:physics/0508025, accessible via http://arxiv.org/abs/physics/0508025)

Huber, J. C. (1998). The underlying process generating Lotka's Law and the statistics of exceedances. *Information Processing & Management*, *34*(4), 471–487.

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129.

Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature*, *415*(6873), 731–732.

Podlubny, I. (2005). *A note on comparison of scientific impact expressed by number of citations in different fields of science*. Technical University of Kosice.