Estimating Research Productivity from a Zero-Truncated Distribution

Timo Koski[1], Erik Sandström[2] and Ulf Sandström[3]

[1] *tjtkoski@kth.se*
Royal Institute of Technology, Dept Mathematical Statistics, SE-100 44 Stockholm, Sweden

[2] *erik@sanskript.se*

[3] *ulf.sandstrom@indek.kth.se*
Royal Institute of Technology, Dept Indek, SE-100 44 Stockholm, Sweden

Introduction

Whilst methods for citation analysis have developed significantly during the last twenty years, the same cannot be said regarding methods for publication productivity analysis. ”Research productivity”, “Scientific productivity” and “Publication productivity” are frequently used keywords in about one and a half thousand ISI-articles over the years, but a closer look reveals little of methodological development with regard to the measurement of “productivity” and few attempts to explicitly contribute to such a development.

This paper will address a fundamental issue in the practical study of scientific productivity, i.e. the calculation of papers per researcher. This task, which at first sight seems quite simple, is often restricted by the properties of the data available. Publication databases, such as Web of Science and Scopus, only contain information on actual (publishing) authors within a given time period, not the full population of publishing and non-publishing “authors”. Hence, a publication frequency distribution based on such data will be zero-truncated; the zero-class (number of non-publishing “authors”) will be missing.

For example, if the productivity of men and women are to be investigated, one might categorize the publishing authors by name and divide the total number of papers by the number of actual authors. However, this calculation will not produce a trustworthy measure of male and female productivity since the proportion of non-publishing (potential) authors might differ between the gender categories. A non-biased measure requires knowledge of the full population, including the number of non-publishing authors (i.e. the zero-class of the publication frequency distribution).

Actual authors and potential authors

“Potential authors” is a concept used by the Budapest group to ex ante denote the total population of researchers that could publish papers, or to ex post denote researchers that could have, but has not, been publishing within a given time period [1; 2]. The (ex post) potential authors include active researchers that have been publishing before or after the given time period but for different reasons have not been publishing during the given time period.

Productivity comparisons between different categories, e.g. field, gender or country requires knowledge about the total population of authors, potential as well actual authors. In a field with low paper productivity, e.g. the social sciences, the proportion of zero-class (the potential authors) will be high in relation to the total population (actual and potential authors). In contrast, in fields with high productivity, such as the medical and natural sciences, the proportion of potential authors will be low. Comparisons based solely on actual authors will thus be misrepresentative.

Actual authors can be extracted from publication databases such as Web of Science and Scopus. Potential authors will, however, not be included in these. An estimate of the number of potential authors would either require detailed data concerning the entire researcher population or the use of statistical estimates, such as the one presented in this paper. A

successful method for estimating the number of potential authors in a productivity distribution (i.e. the zero-class of the publication frequency distribution) would enable the creation of more advanced productivity measures.
The objective of this paper is to contribute to the discussion on productivity from a scientometric perspective. If it is possible to estimate the zero-class of a truncated publication frequency distribution, then it would, in principle, be possible to create advanced (e.g. field normalized) productivity indicators (3).
Hitherto, the most interesting discussion on publication productivity has been given within the framework of frequency distributions [4], which is a core element in bibliometric theory. The Waring distribution is a statistical distribution used for describing publication productivity processes. The distribution was originally introduced by H.A. Simon [5] as a generalization of the Yule distribution and further analyzed by J.O. Irwin in [6], who gave the distribution its current name.

Publication productivity based on frequency distributions

The Waring distribution is a probability distribution p_k on the non-negative integers $k = 0,1,2,\ldots$, and can be defined as follows.

$$p_0 = \frac{\sigma}{\alpha + \sigma} \tag{1}$$

$$p_k = \frac{\alpha + \beta \cdot (k-1)}{\alpha + \beta \cdot k + \sigma} p_{k-1}, \quad k = 1,2,\ldots. \tag{2}$$

The three parameters in (1) and (2) satisfy $\alpha > 0$, $\beta \geq 0$ and $\sigma > 0$. This way of expressing the Waring probabilities reflects the immigration-birth-emigration process introduced by W. Glänzel and A. Schubert in [7] for modelling of scientific productivity. This model was later applied to the citation process in [8]. In the setting of [7] the parameters α and β are coefficients in the linear rates f_i that are proportional to the frequency of authors with i published papers. This means, in plain words, that success breeds success, or, that there is a cumulative advantage in higher levels of productivity. The parameter σ is proportional to the total number of authors and is the rate of external source emitting new authors. There is also a leakage parameter in the model [7], but this does not influence the equilibrium probabilities (= a Waring distribution) of the immigration-birth-emigration process.
Usually the Waring distribution is re-parameterized so that

$$\rho \leftrightarrow \frac{\sigma}{\beta}, \alpha \leftrightarrow \frac{\alpha}{\beta}.$$

Then

$$p_0 = \frac{\rho}{\alpha + \rho} \tag{3}$$

$$p_k = \frac{\alpha + (k-1)}{\alpha + \rho + k} p_{k-1}, \quad k = 1,2,\ldots. \tag{4}$$

Then it can be shown that the expectation μ of the distribution is

$$\mu = \frac{\alpha}{\rho - 1} \quad if \quad \rho > 1. \tag{5}$$

The condition $\rho > 1$ is equivalent to $\sigma > \beta$, which means that the rate of infusion of new authors is higher than the rate of transfer of authors to higher publication numbers.

Let us now suppose that X is a random variable with the Waring distribution with parameters ρ and α, i.e.,

$$\Pr(X = k) = p_k \quad k = 0,1,2,\ldots.$$

with p_ks given in (3) and (4). Then it holds that the truncated expectation of X, $E[X \mid X > k]$, is given as

$$E[X \mid X \geq k] = \mu + (k+1) \cdot \mu_1, k = -1,0,1,\ldots \tag{6}$$

where $\mu (= E[X \mid X > -1])$ is given in (5) and $\mu_1 = \frac{\rho}{\rho - 1}$. This is shown in [9], where it is also shown that the property (7) is a characterization of the Waring distribution, i.e., the Waring distribution has this property, and if (7) is true for a distribution, then it must be a Waring distribution. There is an elegant simplified proof of these facts in [10].

The Waring probability mass function p_k is an example of a power law, since it holds that

$$p_k := P(X = k) : k^{-(1+\rho)}, \quad as \quad k \to \infty. \tag{7}$$

This means that we can call ρ the tail parameter, as it controls the tail of the distribution. A probabilistic analysis that gives (9) as a special case is given in [11].

Power-law tails have been observed in the distributions of the sizes of incomes, cities, internet files, biological taxa, and, after the sequencing of genomes, in (size) distributions of molecular parts, see, e.g., [12; 13].

If it holds exactly that

$$p_k = c \cdot k^{-\gamma}, \tag{8}$$

where c is the normalizing constant, the probability mass function in (8) is known as Lotka's Law, and was found by Lotka (1926) as a bibliometric distribution on the number of authors making contributions. The work by Egghe [14] gives one theoretical justification of the law, which arises in many situations even outside bibliometrics. A similar law had been earlier found by the economist V. Pareto, as a frequency of wealth as a function of income category (above a certain bottom level). In plain words this means: most success seems to migrate to those people or companies who are already popular.

The state probabilities of well known birth-and-death processes are a natural source of power law distributions [12]. This was first realized by G.U. Yule, who established a model (a pure birth process) to explain the observed size distribution of genera with respect to the number of species [15]. Yule obtained a special case of the following probability mass function, which was later generalized in [5]

$$p_k = \delta B(\delta + 1, k), k = 1,2,\ldots,. \tag{9}$$

Here $\delta > 0$ is real, $B(\delta + 1, k)$ is the *Beta function*, i.e.,

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}, \tag{10}$$

where $\Gamma(\cdot)$ is, for z with positive real part, the Euler gamma function.

It can be checked that if X is a random variable with the Waring distribution with parameters ρ and α, then

$$\Pr(X = k \mid X > 0) \to \rho B(\rho + 1, k)$$

as $\alpha \to 0$.

The publication productivity data is by its very definition zero-truncated, i.e., as there is no information of those that are not publishing (in a certain period of time). But as is clear from (7), the Waring distribution is not hampered by the truncation, as pointed out by [4]. In fact the linear expression in gives a way of finding by linear regression against k the estimated intercept $\hat{\mu}$ and this can be used to estimate the frequency of zero via (3) as

$$p_0 = \frac{\alpha + \hat{\mu}}{\alpha(\hat{\mu}+1) + \hat{\mu}} \tag{11}$$

if α is known or estimated from data.

The problem of estimation of zero-frequency (under a Poisson distribution) from zero-truncated data was first considered by A.G.M McKendrick [16]. McKendrick was considering a case of estimating the number of individuals in an Indian village, who were susceptible to infection but did not develop the symptoms. He developed a differential equation and solved it to get the negative binomial distribution, from which he obtained the Poisson distribution as a limiting case. In fact this is related to the Waring distribution, which can be obtained by mixing a negative binomial distribution with a beta distribution with parameters α and ρ. McKendrick developed a moment estimator to find the number of individuals susceptible to infection but did not develop the symptoms. His data contained the number of individuals that did not develop the symptoms, including thus the immune ones. The work of McKendrick and the development of it by J.O. Irwin [17] is surveyed and put into a modern framework in [18]. Burrell points out in [19] that M.G. Kendall (1960) raised the question of estimating the "potentially contributory class" (of journals), when discussing Bradford's work on journal productivity.

4. Testing the Accuracy of Waring Based Estimation

Constructing data sets for precise tests

Earlier empirical studies regarding Waring based estimation of the zero-class have been subject to an important shortcoming: lack of precise data sets. Studies have been performed using truncated distributions, without knowledge of the actual number of zero-frequencies, which has made it difficult of to assess the accuracy of the results.

Precise studies of the accuracy of Waring based estimation of the zero-class (non-publishing "authors") require known publication frequency distributions that include potential authors. The construction of such publication frequency data sets is set out below, using statistics on researchers associated with two different Swedish universities.

The distributions created below are based on selections of (ex ante) "potential authors", i.e. categories of people that we expect to be publishing research papers. A number of these selected people will not have been publishing during the selected time period (but possibly in another time period) and will thus form the zero-class of the publication frequency distributions. It should be noted that the frequency distributions will vary depending on selected categories of people. For example, if we define potential authors as professors at the universities, the number of non-publishing "authors" will likely be small. However, as we include more categories in the potential author definition, the head of the distribution, i.e. the low producing authors, will likely become larger.

Potential authors at two Swedish universities

Employee data from the time period 2005-2007 were obtained from two Swedish universities[19]. A selection of ex ante potential authors was made based on the title of the employments, where professors, researchers and senior lectures were selected. Two selections of potential authors from the respective universities were hereby obtained, containing 729 and 949 employees.

Publication data were downloaded for each ex ante potential author from the Web of Science and compiled into a data set, summarizing the number of publications (article, letter and review) associated with each potential author. In addition, the number of "first author

[19] Linkoping University and Swedish University of Agricultural Sciences

publications" and "reprint publications" (see below) by each author was extracted in listed in the table. Furthermore, a "random author" was selected from each downloaded publication by randomly selecting a single author from the author list of each publication, resulting in a selection of one author per publication. The number of randomly selected authorships of each author was added to the table of publication frequencies.
An author is considered when counting first author publications if listed first among the authors, and for a reprint publication if designated as the corresponding author. Using this paper as an example, Timo Koski would be considered when counting first author publications, Ulf Sandström would be considered when counting reprint author publications, whereas Erik Sandström would only be considered in the all author counting (and possibly the "random author" counting, in which any one of the three authors could be selected).
For each university and authorship type (all, random, first and reprint), a publication frequency distribution, i.e. the number of authors having one publication, two publications and so forth, was compiled (see Table 1).
All zero-frequencies were subsequently removed to construct zero-truncated samples on which the estimations could be performed.

	University 1				University 2			
	All Au	Random Au	First Au	Reprint Au	All Au	Random Au	First Au	Reprint Au
0	[140]	[321]	[321]	[334]	[110]	[370]	[400]	[371]
1	81	159	144	131	172	238	229	224
2	73	95	114	97	117	136	162	130
3	77	53	73	66	102	80	74	86
4	49	32	34	36	77	48	39	57
5	44	29	20	19	59	27	18	32
6	33	12	9	7	46	18	8	10
7	39	8	5	8	40	11	10	9
8	18	5	3	9	34	8	2	9
9	25	5	1	2	25	6	2	5
10	19	3	1	5	25	2	2	2
>10	131	7	4	15	142	5	3	14

Waring based estimation of the population mean

The publication frequency distributions constructed above are zero-truncated samples: the zero-class is missing (or in this case, intentionally hidden). The objective of the method presented and tested in this paper is to estimate this zero-class and in turn the mean of the non-truncated Waring distribution (the population mean).
The method is very simple and carried out using the following steps, as set out in [20].
1. Extraction of the stepwise left truncated sample mean: The mean of the full (zero-truncated) sample was calculated after which all one-frequencies (authors with one publication) were removed (resulting in a one-truncated sample). Subsequently, the one-truncated sample mean was calculated, the two-frequencies removed and so on. The result is a set of data points where the x-axis range from one (zero-truncated) to the maximum value of the distribution, and the y-axis present increasing values (the calculated means).
2. Fitting of straight line: The data points were plotted and a straight line fitted through the points using weighted least square regression. Weights presented in [20] were used. The intercept of the fitted line is an estimate of the mean of the non-truncated Waring distribution.

Simplified Waring estimation

The estimation presented above may be further simplified and calculated only based on the sample mean, the total number of frequencies and the fraction of one-frequencies [2; 21].

Using this method, estimation of the mean of the non-truncated Waring distribution μ can be carried out using the following formula:

$$\mu = \frac{S}{S(1-f)/[x(1-f)+f(1-x)]} \tag{12}$$

where S is the total number of papers, x is the average number of papers per author and f is the fraction of authors with exactly one paper.

Agreement between expected and estimated productivity

For each constructed zero-truncated distribution, the population means were estimated using the Waring approach. The results of the estimations and the actual population means are presented in Table 2.

Table 2. Comparison between actual numbers and Waring estimates

	University 1				University 2			
	All AU	Random AU	First AU	Reprint AU	All AU	Random AU	First AU	Reprint AU
Calculated	6,56	1,55	1,38	1,65	5,43	1,55	1,30	1,68
Estimated	7,26	1,75	1,39	2,04	5,22	1,51	1,17	1,63
(Simplified)	6,98	1,64	1,67	2,03	4,82	1,47	1,35	1,65

The estimated values are for the most part very good. In many cases the estimates are within a 10 % margin from the expected values. Only in a few cases the estimates are considerably far from the expected values (>20 %). The Simplified Waring estimation performs on par with the full version.

Testing the Reliability of the Estimate

The results above indicate that Waring based estimation of the zero-class in general produce good results. A concern is, however, that the estimates will be very sensitive to small variations in the sample, i.e. that the reliability of the estimates will be low. That is certainly the case in the estimates provided above since they are based on relatively small samples (~500-1 000 authors). But if we do not need to know the zero-class, larger samples can easily be created. In the following section, the reliability of the estimates is studied using larger samples.

Field delimited Nordic Publications

The potential author data sets outlined above provide a full population of publication frequencies, including the zero-class. This provides for detailed comparisons of expected and estimated population means. However, the data sets are rather small and the estimates can therefore be expected to be unstable. To study the confidence interval of estimates on a larger sample, further data sets has been compiled.

From the Web of Science, publications with Nordic addresses published between 2003 and 2006 were downloaded. A selection of Nordic authors was obtained by extracting "first authors" (see above) from the downloaded publications and connecting these to the designed addresses. Authors with non-Nordic addresses were removed. The restriction to first authors was necessary since other authors, in the case of collaborative papers, could not be associated with specific addresses.[20]

[20] Complete author-address information only exists in Web of Science after 2008. An alternative solution would have been to use reprint authors (which has revealed similar results).

The names of the selected set of author fractions were manually adjusted to distinguish between homonyms and to harmonize author fractions relating to the same person. The number of first authorships of each distinctive author represented in the data were extracted and compiled into a table.
Each publication was designated to one of seven "fields" based on the classification of ISI subject categories proposed by Zhang et al. [22]. Following this, each distinctive author was designated to the field where the author had most publications. In cases where the number of publications was equal for two fields, one field was randomly selected.
For each field, a publication frequency distribution, i.e. the number of authors having one publication, two publications and so forth, was compiled. The number of potential authors (authors with zero publications) in the selected time period is not known. Hence, the distributions are zero-truncated.

Table 3. Publication frequency distributions for seven fields of science

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
1	1759	2770	3761	2257	6452	1548	2121
2	759	1255	1611	827	2835	647	659
3	385	627	961	368	1593	387	285
4	221	327	516	211	903	172	126
5	112	183	333	108	500	98	52
6	79	107	179	45	327	58	28
7	53	57	149	29	259	40	16
8	24	44	97	17	138	29	7
9	25	23	76	19	113	28	6
10	27	14	44	12	85	12	5
>10	45	37	224	26	352	46	13

[1] Agriculture & Food Science; [2] Biology, Environmental Science & Geography; [3] Chemistry, Physics & Engineering; [4] Computer Science & Mathematics; [5] Life Science & Medical Science; [6] Psychology & Education; [7] Sociology, Economics & Political Science

Bootstrap

In several simple applications of statistical methods to data the uncertainty of an estimate may be assessed by a mathematical calculation based on an assumed probability model for the available data. In more complex situations the mathematical analysis may be time consuming and difficult, and may require additional simplifications that make the results unreliable. Bootstrap is a well established method for obtaining reliable standard errors and confidence intervals for estimates of interest [23]. The main idea is to resample from the original data and thus create replicate data sets from which the variability of the estimate can be inferred without analytic calculation.
With regard to the Waring model of scientific productivity we want basically to estimate the parameter μ, which is the mean of the non-truncated Waring distribution using the linear representation (11) from the above, i.e.,

$$y_s = \mu + s \cdot \mu_1 + e_s, s = 0,1,\ldots,s_{max}$$

where y_s is the left truncated sample mean i.e., an estimate of $E[X \mid X \geq s]$, s_{max} is the maximum value of the publications in data and e_s are respective random deviations (or residuals) of y_s from the 'true' regression line. Then by fitting of straight line by weighted least squares, as described above, we may estimate the intercept μ the regression coefficient μ_1. The question is to obtain a figure of the uncertainty of the estimate μ . There seems to be no immediate analytic procedure for assessment of this uncertainty, as, in

particular, we should perhaps not assume the homoscedasticity and independence of the residuals.
A bootstrap technique for this is to resample, say B times, the authorship data (described above) thus creating a replicate authorship data. For each replicate data set one calculates the regression line obtaining $\mu_1 \ldots . \mu_2$ from which one can calculate the empirical distribution of the estimate of the intercept its bootstrap mean, bootstrap standard deviation and find fractals to compute a bootstrap confidence interval for the intercept.

Confidence intervals of estimated productivity

The bootstrap method was applied to calculate confidence intervals for estimates of papers per researcher in each field set out in table 2. The estimations were carried out using the simplified Waring estimation (12). The results are presented in Figure 1.

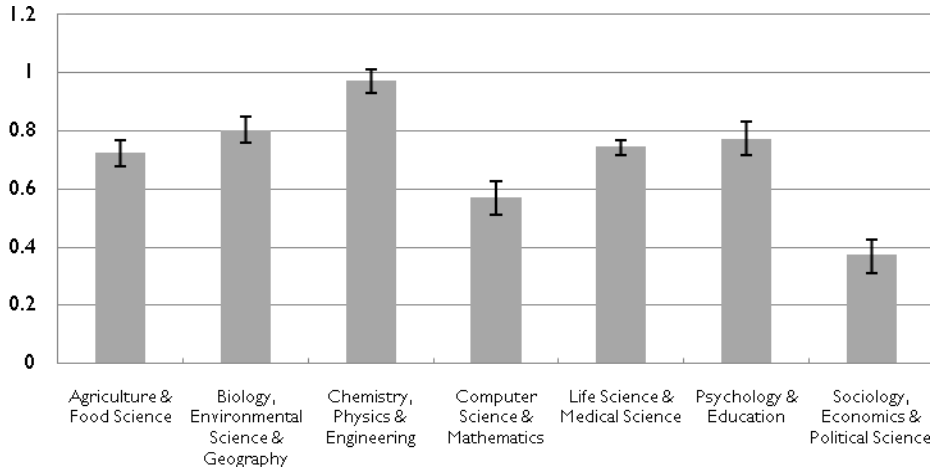

Figure 1. Estimated Papers per Researcher with Confidence Intervals for Seven Fields of Science

The stability of the estimates would be acceptable for a wide range of applications. Except for "Sociology, Economics & Political Science", the confidence intervals are in the range of ±4-8 %. Moreover, the differences between the fields are most often significant. The results further indicate that Waring based estimation of productivity (publications per researcher) can be used for field comparisons, although some additional improvements (e.g. larger data sets and more homogenous field delineations) would certainly be desirable.

Conclusions

The results presented in this paper indicate that Waring based estimation of the zero-class of a zero-truncated publication frequency distribution (and estimates of publications per researcher based thereon) bear real significance and that the estimates are reasonably stable (given a sufficiently large data set). We are convinced that the ability to perform this type of estimates will be the key to more advanced measures of productivity (e.g. field normalized). Consequently, it is our intention to inspire further methodological development on estimates based on publication frequencies. We welcome suggestions of how to improve the Waring model as well as suggestions of alternative methods for mathematical probability.

Acknowledgement-

The authors thank the referees for their valuable comments on an early draft of the paper.

References

A. Telcs & A. Schubert (1986): Publication Potential: an indicator of scientific strength for cross-national comparison. Scientometrics, 9, 5-6, 231-238.
T. Braun, W. Glänzel & A. Schubert (2001): Publication and co-operation patterns of the authors of neuroscience journals. Scientometrics, 51, 3, 499-510.
U. Sandström & E. Sandström (2008): Resurser för citeringar (Resources for Citations (in Swedish). Swedish National Agency for Higher Education, Report 2008:18.
T. Braun, W. Glänzel & A. Schubert (1990): Publication productivity: from frequency distributions to scientometric indicators. Journal of Information Science, 16, 37 44.
H.A. Simon (1955): On a Class of Skew Distribution Functions. Biometrika, 42, 3/4, pp. 425-440.
J.O. Irwin (1963): The place of mathematics in medical and biological statistics. Journal of the Royal Statistical Society. Series A (General), 126, 1-45.
W. Glänzel & A. Schubert (1984): A dynamic look at a class of skew distributions. A model with scientometric applications. Scientometrics, 6, 149-167
W. Glänzel & A. Schubert (1995): Predictive aspects of a stochastic model for citation processes. Information Processing & Management 3, 169-180
W. Glänzel, A. Telcs & A. Schubert (1984): Characterization by truncated moment s and its application to Pearson type systems. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 66, 173-183
C. Dimaki & E. Xekalaki (1996): Towards a unification of certain characterizations by conditional expectations. Annals of the Institute of Statistical Mathematics, 48, 157-168.
W-C. Chen (1980): On the weak form of Zipf's law. Journal of Applied Probability, 17, 611-622.
W.J. Reed, and B.D. Hughes (2002): On the Size Distribution of Live Genera. Journal of Theoretical Biology, 217, 125-135.
W.J. Reed, and B.D. Hughes (2002): From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. Physical Review E, 66, pp. 67-103.
L. Egghe (2005): The Power of Power Laws and an Interpretation of Lotkaian Informetric Systems as Self-Similar Fractals. Journal of the American Society for Information Science and Technology, 56(7), 669- 675.
G.U. Yule (1925): A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S.. Philosophical Transactions Royal Society B, 213, pp. 431-444.
A.G.M. McKendrick (1926): Applications to mathematics in medical problems. Proceedings of Edinburgh Mathematical Society, 44, 98-130.
J.O. Irwin (1959): On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. Biometrics, 15, 324-326.
X.L. Meng (1997): The EM algorithm and medical studies: A historical link. Statistical Methods in Medical Research, 6, 3-23.
Q.L. Burrell (2004): Sample -Size Dependence or Time-Dependence of Statistical Measures in Informetrics? Journal of the American Society for Information Science and Technology, 55(2), 183-184.
A. Telcs, W. Glänzel & A. Schubert (1985): Characterization and statistical test using truncated expectations for a class of skew distributions. Mathematical Social Sciences, 10, 2, 169-178.
Three scientometric etudes on developing countries as a tribute to Moravcsik, Michael. Scientometrics, 23, 1, 3-19.
L. Zhang, W, Glänzel & L.M. Liang (2009): Tracing the role of individual journals in a cross-citation network based on different indicators. Scientometrics, 81 (3), 821-838.
A.C. Davison & D.V. Hinkley (1997): *Bootstrap Methods and their Applications,* Cambridge University Press, Cambridge.