THE MULTI-DIMENSIONALITY OF JOURNAL IMPACT

WOLFGANG GLÄNZEL^{a,b}

^a*K.U. Leuven, Steunpunt O&O Indicatoren, Dept. MSI, Leuven, Belgium*

^b*Hungarian Academy of Sciences, IRPS, Budapest, Hungary*

ABSTRACT. In recent studies the issue of the relatedness between journal impact factors and other measures of journal impact have been raised and discussed from both merely empirical and theoretical perspectives. Models underlying for the underlying citation processes suggest distributions with two or more free-parameters. Proceeding from the relation between the journals' mean citation rate and uncitedness and the assumption of an underlying Generalised Waring Distribution (GWD) model, it is found that the journal impact factor alone does not sufficiently describe a journal's citation impact, while a two-parameter solution appropriately reflects its main characteristics. For the analysis of highly-cited publications an additional model derived from the same GWD is suggested. This approach results in a robust, comprehensible and interpretable solutions that can readily be applied in evaluative bibliometrics.

1 INTRODUCTION

The discussion about use and possible improvement of journal citation indicators is as long as the history of the ISI Impact Factor. Originally used to help select journals for the *Science Citation Index* (Garfield and Sher, 1963), the journal "Impact Factor" (IF) soon evolved to an evaluation tool as well. Garfield himself recognised the power of the IF for journal evaluation and considered it later also a journal performance indicator (e.g., Garfield, 1972). In their review of journal citation measures Todorov and Glänzel (1988) described the role of the impact factor and related journal citation measures as follows.

"Journal citation indicators are commonly used as general measures for various journal characteristics and research impact by different participants in the publication, dissemination, and evaluation process of scientific knowledge. ... Many librarians, information scientists and, sociologists of science already consider journal citation analysis as a practical alternative to subjective judgement. Authors may take citation measures from JCR and use them as possible indicators of journal characteristics. Lists of ranked SCI journals may help potential and real users to identify sources with significant contributions. Editors and publishers may relate high citation impact to a successful editorial practice and policy." (Todorov and Glänzel, 1988, p. 47)

Contrasting with the growing popularity, the widespread and even increasing use of the impact factor, both users and scientists have passed increasing criticism on definition, calculation and actual application practices of this indicator (cf. Moed et al., 1999, Glänzel and Moed,

2002). As a consequence, several attempts have been made to improve the impact factor or to develop supplementary or even alternative journal indicators. The criticism refers to four major issues, particularly, the time windows for which publications and citations are counted, the statistic (using a mean value), the lacking normalisation by the subject and citing journals as well as the possible necessity of supplementing the IF by one or more additional indicators (Glänzel and Moed, 2002). Although the *Institute for Scientific Information* (now *Scientific* [part of Thomson-Reuters], Philadelphia, PA, USA) regularly publishes the ‘Immediacy Index’ along with the Impact Factor in their Journal Citation Reports (JCR), this index has never achieved the same popularity as the impact factor. The introduction of a second, even auxiliary measure, however tells against linearity and tacitly admits the multi-dimensionality of journal citation impact.

Most attempts to relate journal impact to a second statistics are heretofore based on the share of the journal’s uncited papers. Recently van Leeuwen and Moed (2005) have published an empirical study of this matter. They considered uncitedness an additional source of information and subdivided journals, which are covered by the *Science Citation Index* (SCI), into different classes and subclasses according to their output, uncitedness and citation impact. They also found an interesting regularity with regard to the interdependence of uncitedness and impact which is expressed by a strong negative correlation between these two indicators. On the other hand, the correlation between the IF and a high-citation indicator for journals was rather low. Egghe (2008) has taken up the idea that uncitedness of a journal is strongly related to its Impact Factor in order to formulate a unique relation between the two variables resulting in the IF as a function of uncitedness. However, this would tell against the notion of a possible two-dimensionality of citation impact, in general, and of the IF, in particular. This, furthermore, means that a journal’s impact indicator could be calculated more or less directly from the share of its uncited papers.

In order to analyse whether a supplementary indicator would be useful, or even necessary, to characterise a journal’s citation impact, the present paper aims at studying the relationship between impact and uncitedness on the basis of two favoured models in bibliometrics, using the publication output of one complete volume of the SCI database and citation windows ranging between three and 21 years. The main objective of this study is to determine whether (journal) citation impact requires a multi-dimensional approach. If a second indicator proves to yield important additional information indeed, we will answer the question of which regularities underlie the relation between the two indicators and of how a robust and comprehensive estimator can be built for the auxiliary indicator. If there is a solution, it should to be applicable to both informetric theory and evaluative bibliometrics.

2 TWO DISTRIBUTION MODELS FOR CITATIONS

The fact that the uncitedness of two journals might differ while their citation means practically coincide (e.g., Glänzel and Moed, 2002) give empirical evidence that the impact and share of uncited papers are not uniquely determining each other, although they might statistically not be independent either. This leads to the assumption of a distribution model with at least two free parameters. Before the relation between mean citation rate and uncitedness as a function of the parameters of two distribution families will be analysed, we briefly summarise background and basic properties of two prominent scientometric distributions.

2.1 The negative binomial distribution

In scientometrics, practically two two-parameter discrete distributions are in common use. The first one is the NEGATIVE BINOMIAL distribution. This distribution has been early used in the context of measuring the inequality of publication activity and citation impact (Allison, 1980). He introduced the *modified coefficient of variation* as a measure of performance inequality. Assuming an underlying negative binomial distribution, this indicator yields the reciprocal of one of the parameters of the distribution.

The negative binomial distribution can be obtained from the mixture of a *Poisson distribution* with a *Gamma Distribution* (i.e., the parameter of the Poisson distribution is a random variable obeying a Gamma distribution), or results from birth process with linear birth rate. The linearity of birth rate results in a local form of the *cumulative advantage principle* which has, however, little effect on the global properties of the process (cf. Glänzel and Schoepflin, 1994). A third way to obtain this distribution is described by Pólya's urn model (e.g., Johnson and Kotz, 1977). A negative binomial distribution occurs if balls are drawn at random from an urn containing a given number of black and white balls until a certain number of black balls is drawn and each withdrawn ball is immediately replaced. It is clear that the chance of drawing a black or white ball does not change after each withdrawal this way.

The negative binomial distribution is preferably used for data exhibiting overdispersion, that is, in cases when the sample variance exceeds the mean. This is one of the important properties of the negative binomial distribution. It should, however, be stressed that this property is a necessary but not sufficient condition for this distribution.

In one of the first papers capturing the stochastic background of journal citation measures, the standard deviation of the impact factor was calculated using the journal's mean citation rate and the share of uncited papers (Schubert and Glänzel, 1983). This in turn was based on a hybrid ML estimation method for the parameters of a negative binomial distribution (e.g., Johnson and Kotz, 1969). This method uses the mean value and the share of uncited papers. This method has the advantage over the methods of moments that the estimation of the variance of the distribution from the sample variance are usually very unstable because of the long "0-1" chains in the tail of the empirical citation distributions. The fact that this method, which actually used citation mean and uncitedness, functioned properly also supports the assumption of a two-parameter model. The definition of the negative binomial distribution reads as follows.

Definition 1. A non-negative integer-valued random variable X has a negative binomial distribution with parameters N and P if

$$P(X = k) = \binom{N + k - 1}{k} (P + 1)^{-N} \left(\frac{P}{P + 1} \right)^k ; \quad k = 0, 1, 2, \dots \quad (1)$$

where $N > 0$ and $P > 0$.

For the expectation and the probability of the event $\{X = 0\}$ we obtain

$$E(X) = NP \quad \text{and} \quad P(X = 0) = (P + 1)^{-N} . \quad (2)$$

Assuming that the above distribution reflects the citation rate of a paper then $E(X)$ is the expected citation rate and $P(X = 0)$ is the probability of uncitedness. In the following we use the notations

$\mu := E(X)$ and $u := P(X = 0)$. The above mentioned overdispersion property is straightforward, in particular we have $\sigma^2 = E((X - \mu)^2) = NP(P + 1) > NP = \mu$ since $P > 0$. It should be mentioned here that the negative binomial distribution belongs to the domain of attraction of the Gaussian normal distribution.

2.2 The Waring distribution

A second model, the WARING distribution was preferably used for describing publication productivity processes (see Schubert and Glänzel, 1984) but later also in the context of modelling the tail of citation distributions as well (Glänzel and Schubert, 1988a). The Waring distribution can be considered a generalisation of Lotka's Law (see Glänzel and Schubert, 1985) as well.

Definition 2. A non-negative integer-valued random variable X has a Waring distribution with parameters N and α if

$$P(X = k) = \frac{\alpha}{N + \alpha} \cdot \frac{N}{N + \alpha + 1} \cdots \frac{N + k - 1}{N + \alpha + k}; \quad k \geq 0, \quad (3)$$

where $N > 0$ and $\alpha > 0$.

The expectation and the probability of the event $\{X = 0\}$ is obtained as

$$\mu = E(X) = \frac{N}{\alpha - 1}, \text{ provided } \alpha > 1, \text{ and } u = P(X = 0) = \frac{\alpha}{N + \alpha}. \quad (4)$$

In contrast to the negative binomial distribution, the Waring distribution belongs to the class of Paretian distributions with parameter α . Similarly to the negative binomial distribution, the Waring distribution can be obtained from the mixture of a negative binomial distribution with a Pareto distribution (i.e., one parameter of the negative binomial distribution is a random variable which has a Pareto distribution). The tail properties of the Pareto distribution are thus inherited through the mixture. Its parameter α is responsible for the characteristic long tail of this distribution, that disappears as α tends to infinity. Another genesis of the Waring distribution is the limiting distribution of a immigration-birth-emigration process (Schubert and Glänzel, 1984). Schubert and Glänzel introduced a birth-process with linear birth rate, with constant immigration and emigration rates to model the publication activity of scientists in an open system. Due to the effect of immigration, the local form of the *cumulative advantage principle* described in the context of the negative binomial distribution here turns into a global one. This is reflected by the already mentioned characteristic long tail of the distribution. The global cumulative advantage is a consequence of a combination of the linear birth rate with the permanent replenishment of “beginners” through immigration from the external environment. Analogously to the negative binomial distribution, the Waring distribution too can be obtained from Pólya's urn model (e.g., Johnson and Kotz, 1977). In contrast to the negative binomial, here each withdrawn ball is immediately replaced together with a given number of balls with similar colour. This, of course, increases the chance of drawing the same colour again reflecting a further aspect of cumulative advantage.

The necessary and sufficient condition for the existence of a finite expectation is $\alpha > 1$ (cf. Eq. 4), for the existence of a finite variance $\alpha > 2$ since $\sigma^2 = \frac{N\alpha(N+\alpha-1)}{(\alpha-2)(\alpha-1)^2}$. In the latter case the

Waring distribution belongs to the domain of attraction of the Gaussian normal distribution as well.

It is worth mentioning that both the negative binomial and the Waring distribution have a joint special case. The choice of $N = 1$ in the case of a negative binomial distribution and of $N, \alpha \rightarrow +\infty$ if $\frac{N}{\alpha} \rightarrow c > 0$ in the Waring case result in a geometric distribution with parameter P and c , respectively.

Analogously to the overdispersion property of the negative binomial distribution, there is a necessary condition for the Waring distribution as well. This property concerns the relation between the expectation and the probability of the event $\{X = 0\}$. If the random variable X has a Waring distribution then the following inequality holds

$$\frac{1}{u} - 1 = \frac{N + \alpha}{\alpha} - 1 = \frac{N}{\alpha} \leq \frac{N}{\alpha - 1} = E(X) \quad \text{if } \alpha > 1. \quad (5)$$

Equality holds if $N = 0$ or $N, \alpha \rightarrow +\infty$. The first case refers to a degenerate distribution with $P(X = 0) = 1$ implying $E(X) = 0$, whereas the second case corresponds to the geometric distribution. Note that this is not a sufficient condition for obtaining a Waring distribution. This property is, however, not necessary for the negative binomial distribution as readily can be seen. In particular, we can distinguish the following three cases for the negative binomial distribution with $N, P > 0$. According to Bernoulli's inequality (e.g., Mitrović, 1970) we have

$$\mu = NP \begin{cases} < (1 + P)^N - 1 = \frac{1}{u} - 1 & \text{if } N \in (1, +\infty) \\ = (1 + P)^N - 1 = \frac{1}{u} - 1 & \text{if } N = 1 \\ > (1 + P)^N - 1 = \frac{1}{u} - 1 & \text{if } N \in (0, 1) \end{cases} \quad (6)$$

Although the conditions in Eqs. 5 and 6 are not sufficient they can help decide which model may be appropriate. Thus the case $d_{x,u} := \{\mu - (\frac{1}{u} - 1)\} < 0$ can never be obtained from a Waring distribution. In order to illustrate this we analyse the citation distributions of two journals, namely the journal *NATURWISSENSCHAFTEN* and the *AUSTRALIAN JOURNAL OF CHEMISTRY*. We denote the mean value by $\bar{x}$ and the share of uncited papers by $\bar{u}$. Both are unbiased estimators of the expected value $\mu = E(X)$ and the probability $u = P(X = 0)$, respectively. The empirical distributions for the publication year 1980 and the 11-year citation window 1980-1990 are presented in Figure 1. From these samples we can derive the following statistics.

NATURWISSENSCHAFTEN:

$$\bar{x} = 7.48 \text{ and } \bar{u} = 0.193 \Rightarrow \bar{d}_{x,u} = \{\bar{x} - (\frac{1}{\bar{u}} - 1)\} = 3.31 > 0$$

AUSTRALIAN JOURNAL OF CHEMISTRY:

$$\bar{x} = 7.68 \text{ and } \bar{u} = 0.072 \Rightarrow \bar{d}_{x,u} = \{\bar{x} - (\frac{1}{\bar{u}} - 1)\} = -5.28 < 0$$

According to the negative $\bar{d}_{x,u}$ value of the second journal we can conclude that the Waring model is definitely not appropriate for describing the citation distribution of this journal. The negative binomial distribution could be a possible candidate for both journals, however, with $N < 1$ in the first case and $N > 1$ in the second case (cf. Eq. 6). This signifies that distributions with similar means might have different characteristics that are expressed by the range of the parameters. The different shapes of the empirical distributions visualises this property too (see Figure 1).

FIGURE 1. Frequency distribution of citations over papers published in NATURWISSENSCHAFTEN and AUSTRALIAN JOURNAL OF CHEMISTRY in 1980 (citation window: 1980-1990)

3 GENERALISATION

In order to better understand the background of the above inequalities, we extend the analysis to a broader distribution family. The family is called GENERALISED WARING DISTRIBUTION (GWD) and is defined as follows.

Definition 3. A non-negative integer-valued random variable X has a Generalised Waring Distribution (GWD) with parameters α , β and γ if

$$p_k = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + \gamma)}{\Gamma(\alpha)\Gamma(\alpha + \beta + \gamma)} \cdot \frac{\beta\gamma(\beta + 1)(\gamma + 1) \cdot \dots \cdot (\beta + k - 1)(\gamma + k - 1)}{(\alpha + \beta + \gamma)(\alpha + \beta + \gamma + 1) \cdot \dots \cdot (\alpha + \beta + \gamma + k - 1)k!} \quad (7)$$

where p_k denotes the probability $P(X = k)$ for $k = 0, 1, 2, \dots$. $\Gamma(\cdot)$ denotes the gamma function (see e.g., Abramowitz and Stegun, 1964) and α , β and γ are parameters so that the distribution has a sense.

The properties of this distribution family and its applicability to accident theory and cumulative advantage processes have been analysed, among others, by Irwin (1975), Xekalaki (1983), Glänzel et al. (1984) and Burrell (2005). α is a positive real parameter, however, β and γ may be negative and even conjugate complex numbers as well. For their characterisation of the GWD family, Glänzel et al. (1984) introduced the parameter transformation

$$a = \frac{\alpha}{\alpha - 1}, \quad b = \frac{\beta\gamma}{\alpha - 1} \quad \text{and} \quad c = -\frac{(\beta - 1)(\gamma - 1)}{\alpha - 1}. \quad (8)$$

This parameter transformation is used to facilitate the characterisation and analysis on the basis of conditional expectations. Applying the transformation to Eq. 7 results in the following alternative definition of the GWD.

$$p_{k+1} = \frac{(a-1)k^2 + (a+b+c-1)k + b}{(k+1)[(a-1)k + 2a + b + c - 1]} \cdot p_k; \quad k = 0, 1, 2, \dots; \quad a, b, c > 0. \quad (9)$$

The complete description of the distribution family as a function of the (a, b, c) parameter triplet can be found in Glänzel et al. (1984). Figure 2 visualises the GWD domains on the (a, c) plane at a fixed b value (Glänzel, 1994).

The negative binomial distribution occurs if $a = 1$ and the Waring distribution is obtained if $c = 0$ (cf. Fig. 2). We can distinguish the following four cases each.

1. Case ($a = 1$: Negative binomial family)

$$p_{k+1} = p_k \cdot \frac{k + \frac{b}{b+c}}{k+1} \cdot \left(\frac{b+c}{b+c+1} \right); \quad k \geq 0. \quad (10)$$

- 1.1. $c > -b$ corresponds to the NEGATIVE BINOMIAL distribution with $N = \frac{b}{b+c}$ and $P = b+c$.
- 1.2. $c = 0$ corresponds to the GEOMETRIC distribution with $P = b$ ($N = 1$).
- 1.3. $c = -b$ corresponds to the POISSON distribution with $\lambda = b$, i.e.,

$$p_{k+1} = p_k \cdot \frac{b}{k+1}; \quad k \geq 0.$$

- 1.4. $-b > c \geq -b-1$ corresponds to the BINOMIAL distribution with $n = -\frac{b}{b+c}$ and $p = -(b+c)$, provided $\frac{b}{b+c}$ is an integer.

2. Case ($c = 0$: Waring family)

$$p_{k+1} = p_k \cdot \frac{(a-1)k + b}{(a-1)(k+1) + a + b}; \quad k \geq 0. \quad (11)$$

- 2.1. $a > 1$ corresponds to the Waring distribution with $N = \frac{b}{a-1}$, $\alpha = \frac{a}{a-1}$, i.e.,

$$p_{k+1} = p_k \cdot \frac{k + \frac{b}{a-1}}{(k+1) + \frac{a+b}{a-1}}; \quad k \geq 0.$$

- 2.2. $a = 1$ corresponds to the geometric distribution with $P = b$.
- 2.3. $1 > a > 0.5$ corresponds to a finite Waring distribution with $n = \frac{b}{1-a}$, $\omega = \frac{a}{1-a}$, provided $\frac{b}{1-a} \in \mathbb{N}$.
- 2.4. $a = 0.5$ corresponds to the discrete uniform distribution with $n = 2b$, provided $2b \in \mathbb{N}$.

The relation between the expectation and the probability of the event $\{k = 0\}$ will be studied further with the help of a characterisation theorem for the GWD family. This theorem is based on truncated, that is, conditional expectations. It has been proved by Glänzel et al. (1984).

FIGURE 2. The domains of the GWD family on the (a, c) parameter plane at fixed b

Theorem 1 (Glänzel, Telcs, Schubert, 1984). *Let X be a non-negative integer-valued random variable. Then X has a Generalised Waring Distribution (i.e., belongs to Irwin's system) if and only if the following equation holds for all $k \in \mathbb{Z}_0^+$.*

$$E(X|X \geq k) = a \cdot k + b + c \cdot k \cdot E\left(\frac{1}{X+1} | X \geq k\right); \quad k = 0, 1, 2, \dots \quad (12)$$

where $a, b, c > 0$ are positive real parameters so that the distribution has a sense.

A stability theorem for the above characterisation (see Glänzel, 1990, p. 614 *Theorem 2.*) guarantees that Theorem 1 can be applied to empirical distributions as well, in a sense of an “asymptotic” characterisation. Theorem 1 has the further advantage of providing an efficient method for the estimation of the parameters of the GWD. Similarly to the hybrid ML estimation method for the parameters of the negative binomial distribution this method too uses the sample mean as unbiased estimator of the parameter b and the other two parameters are estimated on the basis of a weighted regression according to Eq. 12. In the case of two-parameter subfamilies, such as the Waring and negative binomial family, a pure weighted regression is used for both parameters, that is, the mean is not used in these cases. The efficiency of this method has been studied for the Waring distribution by Telcs et al. (1985). This application is visualised on the Irwin map for the frequency distribution of citations over papers in the journals NATURWISSENSCHAFTEN and AUSTRALIAN JOURNAL OF CHEMISTRY. The programme IrWin, which is briefly described in Glänzel (1994), is used for parameter estimation and visualisation (see Figure 3). The estimated parameters of the journal NATURWISSENSCHAFTEN are $a = 1.00, b = 7.48, c = 3.82$, those of AUSTRALIAN JOURNAL OF CHEMISTRY are $a = 1.20, b = 7.68, c = -3.79$. While the first distribution is a perfect negative binomial distribution, the second one might be well approximated by a negative binomial distribution if

FIGURE 3. The citations distribution of the journals NATURWISSENSCHAFTEN (left) and AUSTRALIAN JOURNAL OF CHEMISTRY (right) on the Irwin map. The location of the journals is marked by black dots.

the number of parameters has to be reduced to two. However, the Waring model assuming $c = 0$ would not work properly for either distribution (cf. Figure 3).

Finally, the characterisation theorem can also be used to derive the theoretical groundwork for the impact–uncitedness relation for empirical distributions. First, we assume that $u \in (0, 1)$, that is $u \neq 0$ and $u \neq 1$. For $k = 1$ we obtain the following expression from Eq. 12 in Theorem 1.

$$\frac{\mu}{1-u} = a + \mu + c \cdot \frac{y_1}{1-u}, \quad (13)$$

where $y_1 := y - u$ and $y := E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty} \frac{p_k}{k+1}$.

Rearranging Eq. 13 results in $u\mu = a(1-u) + cy_1$.

Using the notation $v := (\mu + 1)^{-1}$, we obtain

$$u(\mu + 1) = \frac{u}{v} = a(1-u) + cy_1 + u.$$

Since the l.h.s. of this equation is always positive if $u \in (0, 1)$, we finally have

$$v = \frac{u}{(a-1)(1-u) + cy_1 + 1}. \quad (14)$$

Furthermore, Eq. 14 holds even if $u = 1$ since $P(X = 0) = 1 \Leftrightarrow E(X) = 0$, i.e., $u = 1 \Leftrightarrow v = 1$. Similarly, we have $u = 0 \Leftrightarrow v = 0$, provided the distribution is infinite. Otherwise the denominator of Eq. 14 might vanish and we obtain a non-trivial solution $v \in (0, 1)$ from Eq. 12.

Remark 1. The transformation $x \mapsto v = (x+1)^{-1}$ provides the advantage that both variables u and v take their values in the same finite interval $[0, 1]$.

Remark 2. We can immediately conclude that $v < u$, if $a > 1$ and c is positive, as well as $v > u$, if $a < 1$ and c is negative. Other general cases, e.g., $a > 1$ and c is negative, or $a < 1$ and c is positive, are more complicated and will therefore not be discussed here, however, the cases of the complete negative binomial and Waring families can readily be derived from Eq. 14. We just have to note that $v > u$ or $v < u$ according as the denominator of Eq. 14 is less or greater than 1, i.e., $[(a-1)(1-u) + cy_1]$ is negative or positive, respectively.

The solutions for the negative binomial and the Waring family straightforward. In particular, substitution of $a = 1$ and $c = 0$, respectively, in Eq. 14 results in the following two cases.

1. Case ($a = 1$: Negative binomial family)

$$v = \frac{u}{cy_1 + 1}$$

- 1.1. Negative binomial and $c > 0 \Rightarrow v < u$ for $u \in (0, 1)$.
- 1.2. Geometric distribution, i.e., $c = 0 \Rightarrow v = u$ for $u \in (0, 1)$.
- 1.3. Negative binomial and $0 > c > -b \Rightarrow v > u$ for $u \in (0, 1)$.
- 1.4. Poisson distribution, i.e., $c = -b \Rightarrow v > u$ for $u \in (0, 1)$.
- 1.5. Binomial, i.e., $c \geq -b - 1 \Rightarrow v > u$ for $u \in (0, 1)$.

2. Case ($c = 0$: Waring family)

$$v = \frac{u}{(a-1)(1-u) + 1}$$

- 2.1. Waring, e.g., $a > 1 \Rightarrow v < u$ for $u \in (0, 1)$.
- 2.2. Geometric distribution, i.e., $a = 1 \Rightarrow v = u$ for $u \in (0, 1)$.
- 2.3. Finite Waring, e.g., $0.5 < a < 1 \Rightarrow v > u$ for $u \in (0, 1)$.
- 2.4. Uniform distribution, e.g., $a = 0.5 \Rightarrow v > u$ for $u \in (0, 1)$.

Note that in both cases $v = u$ if $u = 1$ or $u = 0$ (in the latter case, provided that the distribution is infinite). The plots in Figure 4 visualise the (u, v) -relations for selected parameters. Any arbitrary but fixed parameter uniquely determines this relation. The finite members of the corresponding family form the extension of the infinite members in the upper triangle (see binomial and finite Waring distribution in Figure 4). The geometric distribution is represented by the diagonal straight line $v = u$. This line separates the infinite and finite Waring distribution (see right-hand plots in Figure 4). The same function is fulfilled by the Poisson distribution for the negative binomial family (see solid line in the left-hand plots of Figure 4). The possible domains show specific patterns for both families. Figure 5 gives a more detailed picture of the possible domain of the negative binomial distribution. Different N parameter values ranging from 0.1 till $+\infty$ according to the parameter system in Eq. 2 have been selected.

4 THE EMPIRICAL CASE

In this section the theoretical considerations of the previous sections will be applied to a large empirical sample. The citation rates of all journals covered by the Science Citation Index in 1980 have been calculated for different citation windows. In order to obtain reliable estimators and

FIGURE 4. (u, v) -plots for selected Irwin-type distributions with different parameters

indicators, only journals with at least 100 publications of the type *article*, *letter*, *note* and *review* have been selected. Citations have been counted for these document types only. The following citation windows have been chosen for the exercise: three years (1980–1982), six years (1980–1985), eleven years (1980–1990) and 21 years (1980–2000). The variables u and v in Eq. 14 are replaced by the corresponding empirical values $\bar{u}$ (share of uncited papers) and $\bar{v} = \frac{1}{\bar{x}+1}$, where $\bar{x}$ is the journal's impact measure. The use of $\bar{v}$ instead of $\bar{x}$ provides the advantage that outliers need not to be omitted. Both statistics are always calculated for the same citation window of 3, 6, 11 and 21 years, respectively.

The results are presented in Figure 6. The $\bar{v} = \bar{u}$ diagonal represents the geometric distribution. The first observation concerns the flow of dots representing the journals' location in the diagram from the right upper corner in the first diagram to the lower left-hand corner as the citation window grows. This is an obvious consequence of decreasing uncitedness and growing impact with increasing citation window. The second observation concerns the shape of the scatter plot and is in line with the theoretical domains of the negative binomial distribution (cf. Figure 5). Almost exactly 50% of all journals are located below the 'geometric-distribution' line. 81% of all journals do not change their location with respect to this line during the different

FIGURE 5. (u, v) -plot for negative binomial distributions with different parameters N . From left to right: Poisson ($N = \infty$), $N = 5$, $N = 2$, Geometric distribution ($N = 1$), $N = 0.5$, $N = 0.25$, $N = 0.1$

periods. The six-year citation window already provides sufficiently stable results. The following analysis is therefore based on this citation window.

The scatter plots in Figure 6 also substantiate that these shapes are not the consequence of a simple correlatedness; they are in fact determined by distribution-specific properties that can, for instance, in the case of the negative binomial distribution be expressed by the parameter N (see Figure 5). One may therefore expect that this indicator can provide additional information to the mean citation rate. In a first step, this parameter has been estimated for each of the above 1295 journals for the citation window 1980–1985.

In order to obtain a robust parameter estimation method, instead of the weighted regression according to Theorem 1, Eq. 12 with $k = 1$, i.e., Eq. 13 was used. Substituting $c = x(\frac{1}{N} - 1)$ from Eq. 14 in Eq. 13 and rearranging the equation yields the following estimator $\hat{N}$ of the parameter N . We just mention in passing that this method can be considered a weighted regression too, namely with weight 1 for $k = 1$ and 0 each for every $k > 1$.

$$\hat{N} = \frac{\bar{x} \cdot \bar{y} - \bar{x} \cdot \bar{u}}{\bar{x} \cdot \bar{y} + \bar{u} - 1} \text{ and } \hat{P} = \frac{\bar{x}}{\hat{N}} \quad (15)$$

This estimation method has already been successfully applied in the context of a stochastic model for the ageing analyses of scientific literature (cf. Glänzel and Schoepflin, 1994). This model was actually based on a non-homogeneous birth process with double-exponential decay which resulted in a negative binomial citation distribution for any citation window. An important precondition for serving as supplementary indicator to the mean value is uncorrelatedness or at least very low correlation with the empirical mean.

In a second step we have tested the appropriateness of parameter N under this aspect. Of course, the estimators of the two parameters N and P of the negative binomial distribution must

FIGURE 6. $(\bar{u}, \bar{v})$ -plots for 1295 journals with at least 100 publications each in 1980

theoretically be independent. According to the expectation, the correlation with $r^2 = 0.0048$ proved extremely weak indeed. The correlation between the variables $\hat{N}$ and $\bar{x}$ proved to be low as well ($r^2 = 0.1018$), however the Student-test tells against the hypothesis of independence. The t -statistic lies somewhat above the critical value at a confidence level of 0.95. Nevertheless, $\hat{N}$ and $\bar{x}$ cannot be considered interdependent at this level either. Since the distribution is determined by two parameters, the correlation between the third pair $\hat{P}$ and $\bar{x}$ must consequently be stronger. This assumption is confirmed by the distinctly greater correlation coefficient of $r^2 = 0.4772$. It should be mentioned that in all cases the correlation of low-impact journals according to their citation mean proved stronger than that of medium- or high-impact journals. This implies that low-impact journals do, in principle, not need additional indicators.

Figure 7 presents the scatter plots $\hat{N}$ vs. $\bar{x}$ for 1284 out of the set of 1295 journals in three selected fields and all fields combined. Eleven practically uncited journals with $\bar{u} \approx 1$ have been omitted since the right-hand side of Eq. 15 is undefined in such cases. The three selected fields are *mathematics*, *physics* and *neuroscience* according to the Leuven–Budapest subject scheme (Glänzel and Schubert, 2003) have been analysed separately.

Analogously to Figure 6, half the journals are located below the $N = 1$ line (cf. upper left-hand diagram in Figure 7). This area corresponds to the lower triangle in Figure 5. Since this parameter is responsible for the shape of the distribution, these distributions are always extremely left-skewed, that is, the probabilities p_k form a strictly monotonous series with the mode at $k = 0$. The journals above the $N = 1$ line are less skewed, and might have their mode at other k values. The condition of having a mode at $k > 0$ is $\frac{p_k}{p_{k-1}} > 1 > \frac{p_{k+1}}{p_k}$. This condition is equivalent with

FIGURE 7. $(\bar{x}, \hat{N})$ scatter plots for SCI journals with at least 100 papers each published in 1980 in selected fields and in all fields combined (citation window: 1980–1985)

$\frac{\mu}{\mu-k} < N < \frac{\mu}{\mu-(k+1)}$ as can readily be obtained from Eq. 1. This illustrates that distributions with the same expectation μ can have their mode at different $k = 0, 1, 2, 3, \dots$ values according to the parameter N .

Thus the journal NATURWISSENSCHAFTEN (cf. Figure 1) can, for instance, be found below the line ($\hat{N} = 0.81$) whereas the AUSTRALIAN JOURNAL OF CHEMISTRY is located distinctly above the line ($\hat{N} = 1.48$) although both journals have similar impact values for this citation window too (4.78 and 4.68, respectively).

The considerable range of different $\hat{N}$ values at the same citation impact can be observed in all fields, that is, the domains above and below the $N = 1$ separation line or parts thereof are not reserved for specific disciplines; they are occupied by journals of all subject fields.

5 CONCLUSIONS

Earlier studies as well as the present analysis have shown that scientometric distributions, in general and citation distributions of scientific journals, in particular can best be described on the basis of multi-parameter distribution models. One reason for this is, on one hand, the flexible shape of the distribution trunk and the specific tail properties. Two approaches to solve this are the already mentioned GWD and the INVERSE GAUSSIAN-POISSON DISTRIBUTION (IGPD) by Sichel (1992). The main problem of these approaches is that, although they usually result in an excellent fit of the empirical data, the interpretation of three free parameters becomes difficult. Another

problem is of rather technical nature. The application of parameter estimation methods for these distributions to citation distributions of individual journals or paper sets at the large scale proved rather difficult.

The daily practice in evaluative bibliometrics still contradicts the theoretical consideration of modelling and optimum fits. The solution for policy-relevant application would therefore be a compromise. This compromise suggests the use of a comprehensible and robust model that is able to describe the basic features of the distribution in an acceptable manner and in accordance with the shape of its head and trunk. The studies by Glänzel and Schoepflin (1994) and Glänzel and Schubert (1995) have shown that the negative binomial process has all these properties and even allow the prediction of future citation impact. Although uncitedness and mean citation rate might at a first sight serve as potential indicators representing such a two-parameter model, the observation made in Section 3 tells against this assumption. In particular, the correlation between $\bar{u}$ and $\bar{x}$ is strong and the relation between these two indicators is defined by one of two (or more) of the free parameters of the distribution as could be seen in Figure 4. However, the $(\bar{x}, \hat{N})$ indicator pair has the potential to fill this gap. The citation mean is already a standard indicator of journal impact and the correlation between these indicators is low; $\hat{N}$ could therefore be used to provide additional information about the shape of the distribution notably in the case when the “citation impact” of different journals is similar. Both indicators uniquely determine the underlying distribution. Within the framework of the GWD system this solution can be considered kind of “projection on the c -axis” in the Irwin map (of course, not in a strict mathematical sense).

The Waring distribution proved not to have the necessary flexibility for describing journal-citation distributions (cf. Eq. 5). Nevertheless, the assumption of a “tail” parameter for the analysis of the high-end of citations can be useful even if this requires a different model (in a way a “projection on the a -axis”, see Figure 2). According to Glänzel and Schubert (1988a) the observed extreme citation rates are not in line with the distribution model describing the “standard” citation patterns. The same applies to the method of “characteristic scores and scales” (CSS) which is based on the tail analysis (Glänzel and Schubert, 1988b) and the analysis of the properties of the h -index introduced by Hirsch (2005) where the assumption of a Paretian model for citation distributions proved to be useful (e.g., Schubert and Glänzel, 2007, Glänzel, 2008). However, this is already another story. For the general representation of journal impact a two-indicator model is necessary and, as previous studies and the above analysis have shown, for practical application sufficient as well. The somewhat stronger correlation between the two indicators for low-impact journals, however, suggests that in cases of very low citation impact the traditional mean-values based measures provides already sufficient information about the reception of these journals.

Acknowledgement. *The author wishes to thank Dr. András Schubert for his comments and helpful advices.*

REFERENCES

- Abramowitz, M., Stegun, I.A. (1964), *Handbook of mathematical functions*. Dover Publications, New York.
- Allison, P. (1980), Inequality of scientific productivity, *Social Studies of Science*, 10, 163–179.
- Burrell, Q.L. (2005), The use of the generalized Waring process in modelling informetric data. *Scientometrics*, 64 (3), 247–270.

- Egghe, L. (2008), The mathematical relation between the impact factor and the uncitedness factor. *Scientometrics*, 76 (1), 117–123.
- Garfield, E., Sher, I.H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14 (3), 195–201.
- Garfield, E. (1972), Citation analysis as a tool in journal evaluation. *Science*, 178, 471–479.
- Glänzel, W., Telcs, A., Schubert, A. (1984), Characterization by truncated moments and its application to Pearson-type distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66, 1984, 173–183. (Correction: *Probab. Th. Rel. Fields*, 74, 1987, 317.)
- Glänzel, W., Schubert, A. (1985), Price distribution. An exact formulation of Price's "Square Root Law". *Scientometrics*, 7 (3-6), 211–219.
- Glänzel, W., Schubert, A. (1988), *Theoretical and Empirical Studies of the Tail of Scientometric Distributions*. In: L. Egghe, R. Rousseau (Eds.), *Informetrics 87/88*, Elsevier Science Publisher B. V., 75–83.
- Glänzel, W., Schubert, A. (1988), Characteristic Scores and Scales in Assessing Citation Impact. *Journal of Information Science*, 14, 123–127.
- Glänzel, W. (1990), Some consequences of a characterization theorem based on truncated moments. *Statistics*, 21 (4), 613–618.
- Glänzel, W. (1994), *IrWin – A characterization tool for discrete distributions under Windows®*. In: R. Dutter, W. Grossmann (Eds.): *Short Communications in Computational Statistics, Proceedings of COMPSTAT '94*, Vienna (Austria), 22–26 August 1994, 199–200.
- Glänzel, W., Schoepflin, U. (1994), A stochastic model for the ageing analyses of scientific literature, *Scientometrics*, 30 (1), 49–64.
- Glänzel, W., Schubert, A. (1995), Predictive aspects of a stochastic model for citation processes. *Information Processing & Management*, 31 (1), 69–80.
- Glänzel, W., Moed, H.F. (2002), Journal impact measures in bibliometric research. *Scientometrics*, 53 (2), 171–193.
- Glänzel, W., Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56 (3), 357–367.
- Glänzel, W. (2008), On some new bibliometric applications of statistics related to the h-index. *Scientometrics*, 76 (3), forthcoming.
- Hirsch, J.E. (2005), An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America*, 102 (46) 16569–16572. (also available at: [arXiv:physics/0508025](http://arxiv.org/abs/physics/0508025), accessible via <http://arxiv.org/abs/physics/0508025>).
- Irwin, J.O. (1975), The generalized Waring distribution. Part I, II, III. *Journal of the Royal Statistical Society A*, 138, 18–31, 204–227, 374–384.
- Johnson, N.L., Kotz, S. (1969), *Distributions in statistics: Discrete distributions*. Houghton Mifflin, Boston.
- Johnson, N.L., Kotz, S. (1977). *Urn models and their application*. Wiley, New York.
- Mitrinović, D.S. (with Vasić, P.M.) (1970), *Analytic inequalities*. Springer-Verlag, New York – Berlin.
- Moed, H.F., van Leeuwen, Th.N., Reedijk, J.A. (1999), Towards appropriate indicators of journal impact. *Scientometrics*, 46 (3), 575–589.
- Schubert, A., Glänzel, W. (1983), Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, 5 (1), 59–74.
- Schubert A., Glänzel, W. (1984), A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6 (3), 149–167.

- Schubert, A., Glänzel, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), 179–184.
- Sichel, H.S. (1992), Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing & Management*, 28 (1), 5–17.
- Telcs, A., Glänzel, W., Schubert, A. (1985), Characterization and Statistical Test Using Truncated Expectations for a Class of Skew Distributions. *Mathematical Social Sciences*, 10, 169–178.
- Todorov, R., Glänzel, W. (1988), Journal citation measures: A concise review, *Journal of Information Science*, 14 (1), 47–56.
- van Leeuwen, Th.N., Moed, H.F. (2005), Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics*, 63 (2), 357–371.
- Xekalaki, E. (1983), The univariate generalized Waring distribution in relation to accident theory – proneness, spells or contagion. *Biometrics*, 39 (4), 887–895.