Journal of Informetrics 7 (2013) 676–682

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A stochastic approach to the relation between the impact factor and the uncitedness factor

Quentin L. Burrell*

Centre for R&D Monitoring (ECOOM), KU Leuven, Waaistraat 6, Leuven, Belgium

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 12 March 2013
Accepted 13 March 2013
Available online 13 May 2013

Keywords:
Impact factor
Uncitedness factor
Gamma-Poisson process
Success-breeds-success
van Leeuwen–Moed lower bound

ABSTRACT

Empirical analysis of the relationship between the impact factor – as measured by the average number of citations – and the proportion of uncited material in a collection dates back at least to van Leeuwen and Moed (2005) where graphical presentations revealed striking patterns. Recently Hsu and Huang (2012) have proposed a simple functional relationship. Here it is shown that the general features of these observed regularities are predicted by a well-established informetric model which enables us to derive a theoretical van Leeuwen–Moed lower bound. We also question some of the arguments of Hsu and Huang (2012) and Egghe (2013) while various issues raised by Egghe (2008, 2013) are also addressed.

1. Introduction

Hsu and Huang (2012) present sets of data relating to given bodies of work – journal output, and output of a regional group of scientists in a fixed time period – in order to compare the mean number of citations accrued and the proportion of uncited items in a given subsequent time period. They also consider two other fixed bodies of work and examine how the same statistics develop over time. Hence they consider two different types of data sets, one static and one dynamic. What is notable is that they are able to show that in both cases the relationship between mean citations and uncited proportion follows, at least approximately, the same mathematical form. This same sort of relationship had previously been illustrated, in a slightly different way, by van Leeuwen and Moed (2005). In the latter paper, the presentation is purely empirical, while Hsu and Huang (2012) propose a mathematical model. The main difference between the two graphical presentations is that Hsu and Huang (2012) use a linear scale for the mean, van Leeuwen and Moed (2005) adopt a logarithmic scale. Here we illustrate both approaches.

2. The framework

While the impact factor (IF) as usually referred to has a very precise definition, in what follows we will rely on a more flexible formulation, used elsewhere, that suits the current investigation. We imagine a fixed body of work – the publications of a particular scientist, a research school, a geographical grouping of workers in a field, a particular journal, a thematic group of journals – accumulated over a fixed period of time. In a sense we are assuming some sort of homogeneity within the body of work. The interest then is in the citations acquired by this collection in a *subsequent* time period, possibly over an extending period of time.

* Correspondence address: 119 Friary Park, Ballabeg, Isle of Man IM9 4EX, via United Kingdom.
E-mail address: quentinburrell@manx.net

1751-1577/$ – see front matter
http://dx.doi.org/10.1016/j.joi.2013.03.001

To formulate this in terms of a general stochastic model, define X_t = number of citations (to a randomly chosen member of the body of work) by time $t \geq 0$ where 0 denotes the time from which citations are counted. The quantities of interest in the analysis are

$$U = U(t) = P(X_t = 0) = \text{proportion of publications uncited by time } t$$

and

$$\mu = \mu(t) = E[X_t] = \text{mean number of citations per publication by time } t$$

Here μ is referred to as the impact factor (IF), while the proportion of uncited items U is referred to as the uncitedness factor (UF) by Egghe (2013). Note that we have retained the U, μ notation of Egghe (2013) but stressing that these are both time dependent (Hsu and Huang (2012) use a rather than μ). The time dependency is an important consideration. For instance, Hsu and Huang (2012) argue that "intuitively, a journal with a high IF should have a low percentage of uncited articles". However, when one considers the way that citations are accumulated over time, it is clear that, for any fixed collection, the IF will tend to increase with time as more citations are accrued while U will tend to decrease as previously uncited articles eventually become cited and hence it is clear that there should be some sort of time-dependent inverse relation between IF and UF.

It is worth pointing out that there are several possible general scenarios within which this problem can be considered. Because of the context reported by van Leeuwen and Moed (2005) and Hsu and Huang (2012), we assume the following:

Scenario: There is a clearly defined selection of journals/scholars/papers whose *subsequent* citations are accumulated over a fixed or extending time period. In the first instance, we assume that citations are gathered in an unchanging environment although we shall see that this can be relaxed to allow for the issue of ageing/obsolescence or other general changes to citation rates.

Note: An obvious variant on the above is the situation where there is a clearly defined initial selection of journals/scholars/papers whose numbers may be added to in the period of observation and all of whose subsequent citations are accumulated over a fixed or extending time period in an unchanging environment. A particular application is in tracking an author's citations over his/her career where new publications are also being added that can in turn attract citations. Work on this model, including the behaviour of the mean and also the role of the uncited items, is reported in Burrell (2007a, 2007b, 2007c, 2009, 2012, 2013).

3. The simple model and its consequences

We refer to a model as the aim is to describe a mathematical framework for the *mechanics* of the citation process, not merely a mathematical function to fit to observed data. The one to be adopted dates back as far as Greenwood and Yule (1920) and has been used in informetrics at least since Morse (1976) and Burrell and Cane (1982), although these latter applications were in the context of borrowings of library materials. The current application is similar in that we have a well-defined collection of "sources" – here journals/papers – accumulating "items" – citations – over time. The source-item framework is of course very familiar in informetrics. Thus we are considering a dynamic process but also one that is not purely deterministic so that a stochastic element should be built in. (The basic construction is included here for the sake of completeness. Readers familiar with this can happily fast forward.)

Assumption 1. Each source in the collection acquires citations according to a Poisson process of rate $\Lambda \geq 0$.

Assumption 2. The citation acquisition rate varies over the collection according to a Gamma distribution of index $\nu > 0$ and scale parameter $\alpha > 0$. Thus Λ has probability density function (pdf)

$$f_\Lambda(\lambda) = \frac{\alpha^\nu \lambda^{\nu-1}}{\Gamma(\nu)} e^{-\alpha\lambda}$$

for $\lambda > 0$.

If we denote by X_t the number of citations acquired by a typical member of the collection by time $t \geq 0$, where 0 is the start time for the counting of citations, then Assumption 1 can be written as

$$P(X_t) = r | \Lambda = \lambda) = e^{-\lambda t} \frac{(\lambda t)^r}{r!} \text{ for } r = 0, 1, 2, \ldots$$

Then incorporating Assumption 2 yields

$$P(r) = P(X_t = r) = \frac{\Gamma(r+\nu)}{r!\,\Gamma(\nu)} \left(\frac{\alpha}{\alpha+t}\right)^\nu \left(\frac{t}{a+t}\right)^r \text{ for } r = 0, 1, 2, \ldots \tag{1}$$

so that X_t has a negative binomial distribution (NBD) with index ν and parameter $p = p(t) = \alpha/(\alpha + t)$.

Note in particular the case $\nu = 1$ which reduces to X_t having a simple geometric distribution. This restricted version of the model has previously been applied in bibliometrics, such as in Burrell (1980).

The constructed stochastic process $\{X_t;\ t \geq 0\}$ is usually called a gamma-Poisson process, or GPP and, although it has stationary increments, they are not independent so that the correlation structure can be investigated.

Note: The GPP has been extensively used in informetrics, initially in the context of library usage but later in citation distributions, see for instance Burrell and Cane (1982) and Burrell (1985, 1986, 1988, 1990a, 1990b, 2001, 2002a, 2002b, 2003, 2005). It should be pointed out that there are other stochastic mechanisms that can give rise to exactly the same process, see for instance Glanzel and Schubert (1995). It is also worth noting that although in practice one usually finds citation frequency distributions that are decreasing, there is no logical a priori reason why this should be assumed and, for some ranges of parameter values, the NBD is an example of a probability mass function that can have its mode away from zero.

From the above we have

Proposition 1. *With the assumptions of the model the crucial quantities are the UF*

$$U = U(t) = P(X_t = 0) = p(t)^{\nu} = \left(\frac{\alpha}{\alpha + t}\right)^{\nu} \tag{2}$$

and the IF

$$\mu = \mu(t) = E[X_t] = \frac{\nu(1 - p(t))}{p(t)} = \nu\frac{t}{\alpha} \tag{3}$$

Proof. The first is just (1) with $r = 0$ while the second is a standard result for the NBD.

Of course, U is restricted to the unit interval while μ can be any non-negative real number. Note that $U(t)$ decreases with time while the mean increases proportional to time. The issue of interest is the relationship between μ and U. Ignoring the time dependence, we have

Proposition 2.

$$U = \left(\frac{1}{1 + \mu/\nu}\right)^{\nu} \tag{4}$$

$$\mu = \nu\left(\frac{1}{U^{1/\nu}} - 1\right) \tag{5}$$

Proof. Essentially, this is just a matter of eliminating t from (2) and (3) of Proposition 1. $U = p_{\nu}$ so that $p = U^{1/\nu}$ while

$$\mu = \nu\left(\frac{1 - p}{p}\right) = \nu\left(\frac{1}{p} - 1\right)$$

and hence (5) follows. Similarly, or by rearrangement, we get (4).

Corollary. *In the limit as $\nu \to \infty$ we have $U = e^{-\mu}$ or $\mu = -\log U$.*

(Here and throughout, logs are taken to be natural logarithms.)

Proof. From (4)

$$U = \left(1 + \frac{\mu}{\nu}\right)^{-\nu} = \left[\left(1 + \frac{\mu}{\nu}\right)^{-\nu/\mu}\right]^{\mu} \underset{\nu \to \infty}{\longrightarrow} [e^{-1}]^{\mu} = e^{-\mu} \tag{6}$$

and hence

$$\mu = -\log U \tag{7}$$

Note: The model assumed by Hsu and Huang (2012) and Egghe (2013) corresponds to (4) in the case $\nu = 1$, in turn equivalent to X_t having a geometric distribution although it is not identified as such by those authors. Hsu and Huang (2012) argue that the distribution arises as a consequence of a cumulative-advantage mechanism although their version of the geometric distribution includes no time parameter. (We will return to this later.) The case given by (6) is derived by Hsu and Huang (2012) as resulting from what they refer to as the random-selection mechanism.

The relationship given in Proposition 2 is illustrated in Fig. 1 for various values of the index ν together with the limiting case given by (7), which we call the boundary case.

Note that the case $\nu = 1$ corresponds to the solid grey line and the boundary case to the dashed grey line given in Figs. 1–3 of Hsu and Huang (2012).

Remark. The above result says that, assuming the GPP model, given the value of the index ν the point (U, μ) lies on the curve given by (5) with the exact point depending on the value of p which in turn depends on both t and α. If instead we had used the same model but incorporating obsolescence as in Burrell (2003) we would again find a NBD for X_t with the same index ν but with a different time dependent form for p, see Burrell (2002b). Hence we would still find the same functional

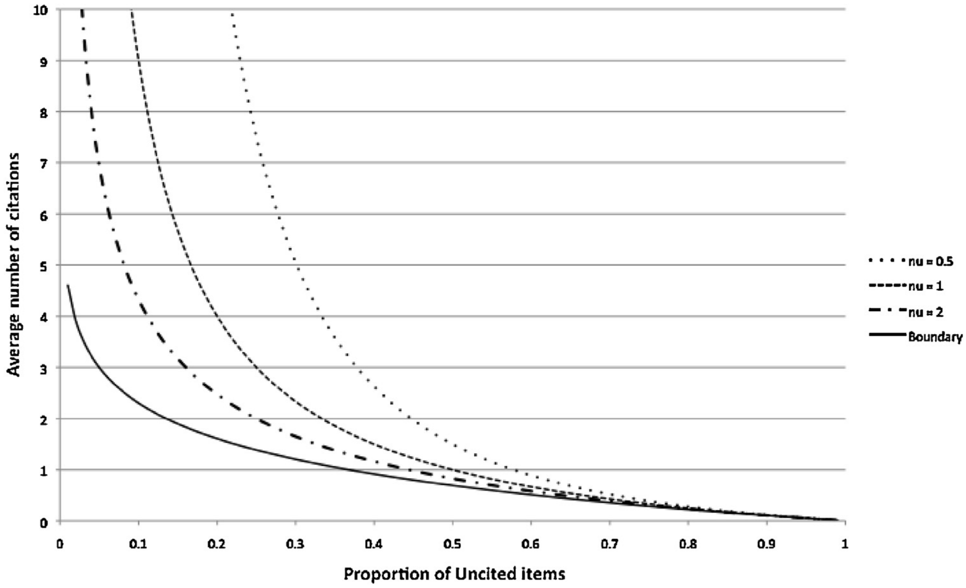

Fig. 1. The relation between the proportion of uncited articles and the average number of citations per article.

relationship between the IF and the UF as in (5). Thus we would expect to find the same sorts of empirical relationships whether or not obsolescence is included in the original stochastic process model.

It is interesting to note that if the development of the collection is observed over an increasing period of time, as Hsu and Huang (2012) report in their study of the Chung Yuan Christian University and the journal Physica A, the GPP model predicts that, although each of U and μ will change with time, the pair (U, μ) will lie on the same curve – given by (5) – since the index ν remains the same. Observed discrepancies are then due to statistical fluctuation. This is in accordance with what Hsu and Huang (2012) illustrate in their Fig. 3 although as noted, their model is not time-dependent. Contrast their Fig. 3 with what is plotted in their Fig. 1, a scatter diagram for 72 different physics journal. The important point is that, according to the GPP model, in the latter case we have two different sources of variation – not just statistical random deviations from a single model but also that for different journals the GPPs may have different values of the index ν.

4. The van Leeuwen and Moed findings

Turning to the situation reported by van Leeuwen and Moed (2005), they considered the relationship between the logarithm of the IF and the UF for journals in all fields covered by the Science Citation Index (SCI) for 1995. (As noted in the Section 1, using a log rather than a linear scale for IF is what distinguishes the van Leeuwen and Moed approach from that of Hsu and Huang.) The scatter plot of the log of IF against the UF (expressed as a percentage) given in their Fig. 1 (reproduced as Fig. 1 of Egghe (2008) and Fig. 3 of Egghe (2010)) is indeed striking. Far from being a random scatter of points, the plot shows that the points seem to coalesce around a clearly defined curve. In their Fig. 2(a), (b) and (c), van Leeuwen and Moed (2005) show similar patterns when restricting the data to specific SCI subject fields. The distinct shape of the curve is fairly described by Egghe (2008) as a horizontal S-shape – first convex, then concave. It is only fair to point out that Hsu and Huang (2012) also considered the effect of a logarithmic transformation for the IF as illustrated in their Fig. 4.

In fact, using the GPP model, we find from (5) the relationship between the log of the IF and the UF is given by

$$y = \log \mu = \log \nu + \log(1 - U^{1/\nu}) - \log(V^{1/\nu}) \tag{8}$$

The shape of the curve is given in the following:

Proposition 3. (i) $\lim_{U\to 0} y = \infty$; $\lim_{U\to 1} y = -\infty$; (ii) $\frac{dy}{dU} < 0$ for all U; (iii) $\lim_{U\to 0} \frac{dy}{dU} = \lim_{U\to 1} \frac{dy}{dU} = -\infty$;

(iv) The curve is initially convex, becoming concave, with a point of inflection at

$$U = \left(\frac{\nu}{1+\nu}\right)^{\nu},\ y = 0.$$

Proof. The first is immediate from (8). The rest all follow from simple calculus. Firstly, differentiating (8) with respect to U and, purely to simplify notation, writing $\rho = 1/\nu$ we find

$$\frac{dy}{dU} = -\frac{\rho U^{\rho-1}}{1-U^{\rho}} - \frac{\rho U^{\rho-1}}{U^{\rho}} = -\frac{\rho}{U(1-U^{\rho})} \tag{9}$$

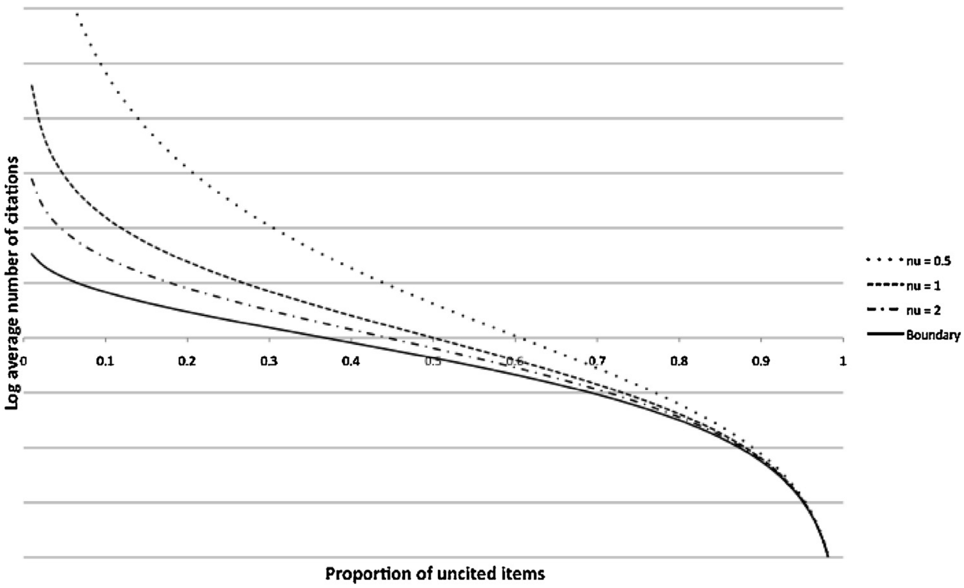

Fig. 2. The relation between the proportion of uncited articles and the logarithm of the average number of citations per article.

Thus (ii) and (iii) follow immediately. For (iv), differentiate (9) once more with respect to U and find, after some simplification, that

$$\frac{d^2y}{dU^2} = \rho\left[\frac{1-(\rho+1)U^{\rho}}{U^2(1-U^{\rho})^2}\right]$$

Hence the sign of the expression is determined by the sign of $1-(\rho+1)\,U^{\rho}$ or

$$1-\left(\frac{1+\nu}{\nu}\right)U^{1/\nu}$$

and, after rearrangement, (iv) follows. The resultant value for μ, and hence y, comes from (4).

Note: As an aside regarding the point of inflection, note that it occurs where the curve crosses the U-axis and also that the U-value decreases as ν increases. In the limit we have

$$\lim_{\nu\to\infty}\left(\frac{\nu}{1+\nu}\right)^{\nu} = \lim_{\nu\to\infty}\left(1+\frac{1}{\nu}\right)^{-\nu} = e^{-1} = 0.3678\ldots$$

An additional aspect of the van Leeuwen–Moed plot that seems particularly striking – to this author, at least – although it does not seem to have raised comment elsewhere, is the very sharp lower boundary of the scatter plot. Because it is so striking- and unexpected – we call this the van Leeuwen–Moed boundary. The following is essentially a reworking of the Corollary to Proposition 2.

Proposition 4. The van Leeuwen–Moed boundary is given by

$$y = \log\mu = \log(-\log U)$$

Proof. The relationships between U and μ in the lim it $\nu\to\infty$ are given in the Corollary to Proposition 2 and the equation for the van Leeuwen–Moed boundary then follows from (7).

To illustrate the results of the Proposition, plots of Eq. (8) are given in Fig. 2 for various values of ν, together with the van Leeuwen–Moed boundary.

What is particularly interesting in Fig. 2 is not just the general way that the graphs reflect the empirical data plotted in Fig. 1 of van Leeuwen and Moed (2005), but the way that the curves funnel in as we move from left to right just as is seen in the original empirical plot. Observe also the behaviour of the point of inflection as remarked on above.

A crucial aspect when considering the application the GPP model to many collections (of articles) is that we have two distinct sources of variation – not just statistical random deviations from a completely specified model but also that for different collections the GPPs - and hence the IF–UF relation – may have different values of the index ν. Thus in graphical presentations such as that of van Leeuwen and Moed we do not just have random scatters of points around a particular curve, we have scatters around many different curves, all superimposed. We believe that this at least in part provides an explanation to the open problem posed at the end of Egghe (2010) regarding the "thicker" part of the van Leeuwen–Moed graph. (Although Egghe (2008) claims to provide a proof of this pattern in general, in fact he only illustrates the relationship for a simple example.)

5. On "success breeds success" and the Hsu-Huang argument

In their presentation, Hsu and Huang (2012) talk of "cumulative advantage" and "the Matthew effect" or "the rich get richer" and use this idea to justify consideration of the requirement that

$$\frac{dP(n)}{dn} \propto P(n) \tag{10}$$

where "the time dependence has been removed from the asymptotic limit".

This seems to be a strange approach. Surely, *success breeds success* – or any of its other expressions – is necessarily a *temporal* phenomenon reflecting the way that the number of citations changes with time – the more citations an item has received by any particular time, the more citations it might be expected to receive in the future? There should also be a probabilistic element – that an item has received a lot of citations does not guarantee its future success, only that it has a good chance of future success. It is thus hard to see how the model given in (10), where time has been removed, can be described as reflecting cumulative advantage.

Eq. (10) says first of all that the number of papers/journals/etc. receiving n citations decreases with increasing n. So how does this reflect success breed success? Furthermore, the rate of decrease is assumed to be directly proportional to the number of sources with n citations, without any supporting argument. In addition, note that in (8), n is a an integer variable and hence the proposed dependency should be through a difference equation, possibly

$$P(n+1) - P(n) = cP(n), \quad \text{where } c < 0$$

from which the geometric distribution follows almost immediately.

In fact, it is the resultant geometric distribution – or rather the simple relation between the geometric parameter $p = P(0)$ and the mean μ – on which Hsu and Huang (2012) base all of their analysis. Egghe (2013) reproduces the Hsu and Huang proof, again justifying the derivation by reference to the Matthew effect but in addition provides what is claimed to be a (heuristic) proof of (1) with $\nu = 1$, without reliance on (10). However, this "proof" relies on "*assuming that the fraction of (the number of articles) representing 1 citation is also the fraction of articles with 1 citation*". At best the argument seems to be circular!

As noted above, a model for "success-breeds-success" must surely include the time dimension and hence be a dynamic model as well as acknowledging that actual numbers of future citations are uncertain. One has only to look at the original presentation of Derek de Solla Price to appreciate his insight (Price, 1976). In the current context, the GPP, with or without ageing or obsolescence, has been shown to have a form of cumulative advantage as a consequence of its correlation structure. The crucial mathematical feature is that, as remarked earlier, the stochastic process $\{X_t;\ t \geq 0\}$ does not have independent increments. The essence of the result is that, given $X_t = r$, i.e. given that the number of citations received by time t is r, the number of *additional* citations by time $t + s$ has a NBD of index $\nu + r$ so that the *expected* number of additional citations is proportional to $\nu + r$. Indeed, the expected number of future additional citations in a given time is a linearly increasing function of the current number of citations. Thus, the more citations an item has received by time t, the more it might expect to receive in a subsequent time period. Success breeds success!

(The mathematical details, including the situation where there is assumed to be obsolescence, can be found in such as Burrell and Cane (1982), and Burrell (1986, 1988, 1990b, 2003).)

6. Concluding remarks

It has been shown that the well-known GPP model gives a reasonable qualitative description of the empirical phenomena reported by van Leeuwen and Moed (2005) and Hsu and Huang (2012). The model is not claimed to be perfect, as has often been acknowledged in the previously mentioned literature, but often does give a good descriptive picture of the empirical phenomena being considered. We believe that we have resolved to a large extent the open problem posed by Egghe (2008) in providing a realistic model incorporating exactly the uncitedness factor. The model also provides further insight into the empirical findings by including the extra parameter ν to model a further source of variation. What we have not done is to give a practical interpretation of the ν parameter. One of the referees asks whether it is perhaps related to characteristics of different disciplines, subfields or journals? This is perhaps the crucial question and one that we invite others to explore through further empirical studies.

Acknowledgements

The author is grateful to the two anonymous referees who provided many constructive comments as well as tempering some of wording in the original and encouraging discussion of possible future work.

References

Burrell, Q. L. (1980). A simple stochastic model for library loans. *Journal of Documentation*, *36*(2), 115–132.
Burrell, Q. L. (1985). A note on ageing in a library circulation model. *Journal of Documentation*, *41*, 100–115.
Burrell, Q. L. (1986). A second note on ageing in a library circulation model: The correlation structure. *Journal of Documentation*, *42*, 114–128.
Burrell, Q. L. (1988). Predictive aspects of some bibliometric processes. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88: Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval* (pp. 43–64). Amsterdam: Elsevier.
Burrell, Q. L. (1990a). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science*, *41*(3), 164–170.
Burrell, Q. L. (1990b). Correlation structure of library circulation data: A case study. Part 2 – Theoretical aspects. *Library Science*, *27*, 237–252.
Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, *52*, 3–12.
Burrell, Q. L. (2002a). On the *n*th-citation distribution and obsolescence. *Scientometrics*, *53*, 309–323.
Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*, 232–235.
Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, *54*(5), 372–378.
Burrell, Q. L. (2005). Are "Sleeping beauties" to be expected? *Scientometrics*, *65*(3), 381–389.
Burrell, Q. L. (2007a). Hirsch's *h*-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25.
Burrell, Q. L. (2007b). On the *h*-index, the size of the Hirsch core and Jin's *A*-index. *Journal of Informetrics*, *1*(2), 170–177.
Burrell, Q. L. (2007c). On Hirsch's *h*, Egghe's *g* and Kosmulski's *h*(2). *Scientometrics*, *79*(1), 79–91.
Burrell, Q. L. (2009). The publication/citation process at the micro level: A case study. *Journal of Scientometrics and Information Management*, *3*(1), 71-77.
Burrell, Q. L. (2012). Alternative thoughts on uncitedness. *Journal of the American Society for Information Science and Technology*, *63*(7), 1466–1470.
Burrell, Q. L. (2013). *The individual author's publication-citation process: a stochastic model*. (In preparation).
Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data (With discussion). *Journal of the Royal Statistical Society (A)*, *145*(4), 439–471.
Egghe, L. (2008). The mathematical relation between the impact factor and the uncitedness factor. *Scientometrics*, *76*(1), 117–123.
Egghe, L. (2010). The distribution of the uncitedness factor and its functional relation with the impact factor. *Scientometrics*, *83*, 689–695.
Egghe, L. (2013). The functional relation between the impact factor and the uncitedness factor revisited. *Journal of Informetrics*, *7*, 183–189.
Glanzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management*, *31*, 69–80.
Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distributions of multiple happenings etc. *Journal of the Royal Statistical Society (A)*, *83*, 255–279.
Hsu, J.-W., & Huang, D.-W. (2012). A scaling between impact factor and uncitedness. *Physica A*, *391*, 2129–2134.
Morse, P. M. (1976). Demand for library materials: An exercise in probability analysis. *Collection Management*, *1*, 47–78.
Price, D. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*, 293–306.
van Leeuwen, T. 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.