

The publication/citation process at the micro level: A case study

Quentin L. Burrell*

September 12, 2008

Abstract

The definitions of h-type indexes, seeking to give a single-number measure of an author's impact, usually involve both the author's productivity in terms of the number of papers published and of the number of citations subsequently received. In studying the evolution of such indexes over time, one therefore needs to consider both the publication and citation processes.

Here we consider the career (so far) of a single scientist so far as his publication and citation records are concerned using data gathered from the *Web of Science* (WoS). Our interest is, in particular, focussed on individual citation patterns as well as the overall cumulation of publications and citations and the distribution of citation rates. We find that the development of the citation process as a whole conforms well to the form speculated by Hirsch (2005, 2007) as well as Burrell's (2007a,b) stochastic model. However, the citation process at the level of individual papers shows discrepancies from the model assumptions and we suggest some possible reasons for this.

1 Introduction

In the paper in which he first presented the h-index, Hirsch (2005) gave a heuristic and deterministic argument to describe the evolution of a scientist's publication/citation career. Burrell (2007) subsequently presented a model which sought to incorporate both the publication and

citation processes in a stochastic formulation. Numerical investigations using Burrell's (2007) model produced results that supported Hirsch's (2005) conjectures.

2 The case study

B. W. (Bernard) Silverman is an eminent British statistician. He gained early recognition as a profound and influential researcher in theoretical statistics, most notably in the areas of density estimation, smoothing techniques and the use of wavelets. He has received many honours and awards, including *Guy Medals* in both Bronze (1984) and Silver (1995) from the *Royal Statistical Society*, and was elected a Fellow of the Royal Society (FRS) in 1997. In this study, we consider Silverman's published papers classed as "articles" by *Web of Science* (WoS), of which there are 62 published between 1976 and 2007, the period covered here.

3 The basic evolution

3.1 The publication process

The simple model for the publication process of an individual scientist proposed by Hirsch (2005) was to assume that the scientist produces papers at a constant rate. In this deterministic model, therefore, a plot of the total number of papers against time will be strictly linear, the slope being the assumed rate. Burrell (2007) in-

* Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB, via United Kingdom
q dot burrell at ibs dot ac dot im

stead proposed that the author produces papers in some random fashion over time. More precisely he proposed that the publication process can be modelled by a Poisson process of rate θ . Thus from the start of his/her publishing career (at time zero), by the end of year n the number of publications Y_n has the distribution

$$P(Y_T = r) = e^{-\theta T} \frac{(\theta T)^r}{r!}, r = 0, 1, 2, \dots$$

and the expected or mean number is given by $E[Y_n] = \theta n$. Note that the parameter θ gives the mean number of publications per year for this author, called the *publication rate*. With this model the actual number of publications does not increase in a strict linear fashion but increases linearly “on average”.

Note that the original formulation was in continuous time while for the empirical data, we are relying on end-of-year data taken from WoS. Hence in the above $n = 0, 1, 2, \dots$. In the case of Bernard Silverman, his first paper appeared during 1976 so we take $n = 0$ to be the end of 1975, $n = 1$ to be the end of 1976 and so on to $n = 32$ being the end of 2007. Looking at the Silverman data, we see from Figure 1 that the data points increase approximately linearly. Indeed, for the least squares (LS) regression line constrained to pass through the origin, as it must, we find $R^2 > 0.99$ and the LS estimate of the rate is

$$\hat{\theta}_{LS} = 2.06.$$

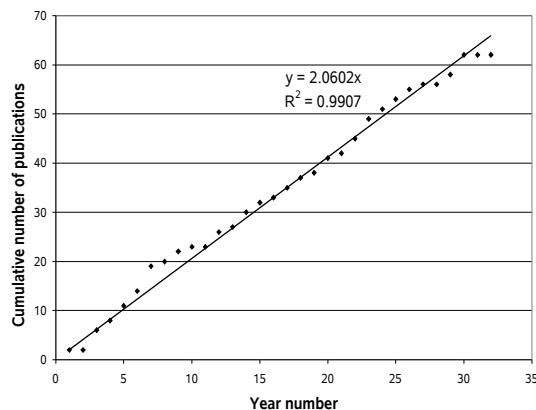


Figure 1: Year-by-year cumulation of Silverman’s published articles

Thus, this example suggests that the assumption of a constant mean rate of publication is tenable.

3.2 The citation process

Hirsch (2005) assumed, for his heuristic, deterministic model, that not only does an author publish at a constant rate but that each published paper subsequently accrues citations at a constant rate, the same for all papers. Thus each paper published in year j receives c (say) citations in years $j+1, j+2, \dots$. It is then easy to show that the total number of citations received by all publications in years $1, 2, \dots, n$ is, by time $n+1$, proportional to $n(n+1)$. We have already seen that Burrell’s (2007) continuous time stochastic model assumes that the publication process is random. For the citation process, he assumes variation occurring in two ways: firstly that each paper receives citations in a random fashion and, secondly, that the (mean) rate at which citations are received varies from paper to paper. It is shown in Burrell (2007b) that, given the precise model assumptions, the (expected) total number of citations received by the end of year n is proportional to n^2 . (The difference in the two results arises in part from the fact that in the discrete time deterministic model a paper can only receive citations in the next year, in the continuous time stochastic model citations can be received at any time after publication.)

Writing $C(n)$ for the cumulated number of citations, the stochastic model assumes that the approximate relationship is $C(n) = \alpha n^2$, i.e. a pure square law. The LS regression estimate of the coefficient α is given by

$$\hat{\alpha} = \frac{C(n)n^2}{n^4}$$

For the Silverman data we find $\alpha = 2.76$ and this gives the fitted LS regression as in Figure 2. Clearly, the fit of the pure square law is not very satisfactory, describing the development only in the most general terms.

In Figure 1 of Hirsch (2007) are shown the cumulated citation distributions with fitted quadratic curves, though showing rather fewer data points than we have here, for two eminent

physicists. The fits look to be quite good although the precise form of the assumed quadratic is not stated nor is any goodness of fit reported. If we relax the strict square law form we see in Figure 3 that, for the Silverman data, the LS quadratic fit constrained to pass through the origin, i.e. the form $C(n) = \alpha n^2 + \beta n$, is excellent with $R^2 = 0.9977$.

However, modelling informetric data should not just be a curve-fitting exercise – one should look at the explanatory details of the model.

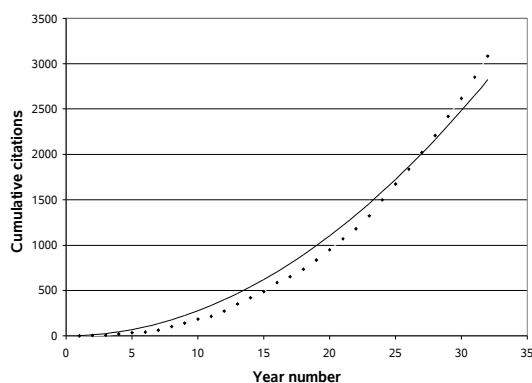


Figure 2. Year-by-year cumulation of citations to Silverman's published articles with fitted pure square curve.

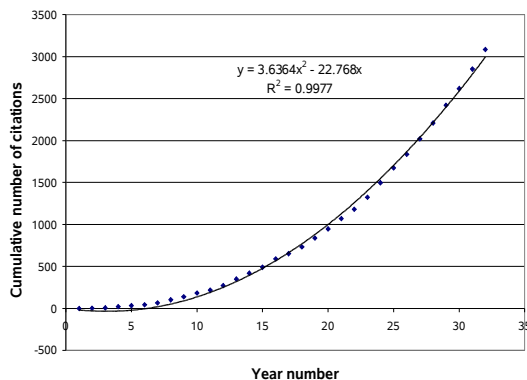


Figure 3. Year-by-year cumulation of citations to Silverman's published articles with fitted (restricted) quadratic curve.

4 Model assumptions

We have seen that the cumulative citation distribution is not quite in accord with the stochastic

model prediction. To see why this might happen we need to look more closely at the basic model assumptions, of which there are two. The first is that any particular publication attracts citations at a constant rate on average – more precisely, according to a Poisson process. The second is that different papers have different rates – more precisely, with rates following a gamma distribution.

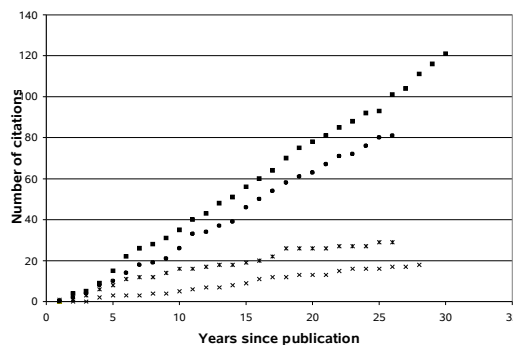


Figure 4(a). Constant citation rate.

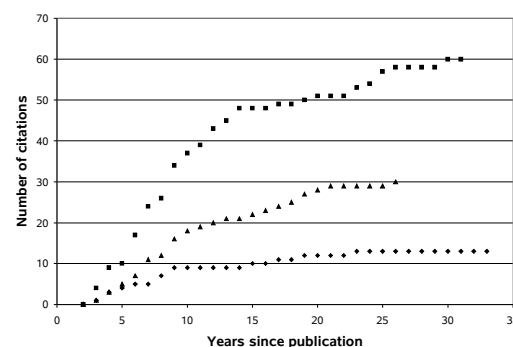


Figure 4(b). Decreasing citation rate.

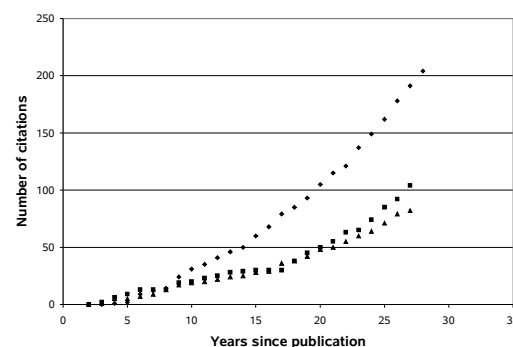


Figure 4(c). Increasing citation rate.

The reason why these specific assumptions were made was that on the one hand they are intuitively reasonable but also that they enable a full analytic form of the mathematical model to be derived. Here we just consider the general assertions using simple graphical methods.

Figures 4(a)–(c) show the cumulative citations for a selection of Silverman’s papers. For each paper we take year 1 to correspond to the year of publication so that the lengths of the graphs vary depending on the year of publication. We have sorted the plots into three groups representing the typical general shapes we have found, though there are exceptions.

In 4(a), where we see the cumulated citations increasing linearly ($R^2 > 0.95$ in all cases), this corresponds to a constant mean citation rate, as assumed in the model. The approximately concave forms as illustrated in 4(b) correspond to papers having a decreasing citation rate. Note that it is often assumed in the literature that papers do suffer from ageing or obsolescence, perhaps because they are superseded by later work, which would lead to a decreasing citation rate. Finally, in 4(c), we see instances approximately convex forms, corresponding to papers with an increasing citation rate. These are, perhaps, papers whose importance is only gradually realised and which become ever more influential as time progresses. Extreme forms of this phenomenon have been referred to as “Sleeping Beauties” (van Raan 2004, Burrell 2005).

Note also that the three sets of graphs all show differences in citation rates, as expected, and that the different forms are not associated with particular sorts of citation rates.

5 Discussion

We have seen that, at least in the case of Bernard Silverman, an author’s publications increase approximately linearly over time; the number of received citations increases approximately quadrat-

ically. These results are in line with model assumptions of Hirsch (2005, 2007) and Burrell (2007a, b). However, at the level of individual papers, there are discrepancies between the observed citation behaviour and that assumed in the stochastic model. In particular the assumption of a constant citation rate for each paper does not hold in many cases. It is remarkable that such discrepancies do not affect the overall behaviour of the citation process in general terms. Why this might be warrants further investigation. Despite these reservations, the stochastic model is the most successful so far presented for the publication/citation process. We suggest that more, similar, work at the micro level could well be fruitful, perhaps leading to comparative studies between different workers in the same field and between workers in different fields.

References

- 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.
- 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.
- Hirsch, J. E. (2007). Does the h-index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), 19193–19198.
- Van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, 59, 467–472.