Stopped Sum Models for Citation Data

Wan Jing Low, Paul Wilson and Mike Thelwall

W.J.Low@wlv.ac.uk, PaulJWilson@wlv.ac.uk, M.Thelwall@wlv.ac.uk
Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of
Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK)

Abstract
It is important to identify the most appropriate statistical model for citation data in order to maximise the power of future analyses as well as to shed light on the processes that drive citations. This article assesses stopped sum models and compares them with two previously used models, the discretised lognormal and negative binomial distributions using the Akaike Information Criterion (AIC). Based upon data from 20 Scopus categories, some of the stopped sum models had lower AIC values than the discretised lognormal models, which were otherwise the best. However, very large standard errors were produced for some of these stopped sum models, indicating the imprecision of the estimates and the impracticality of the approach. Hence, although stopped sum models show some promise for citation analysis, they are only recommended when they fit better than the alternatives and have manageable standard errors. Nevertheless, their good fit to citation data gives evidence that two different, but related, processes drive citations.

Conference Topic
Citation and co-citation analysis

Introduction

Fitting statistical models to citation data is useful both to understand the citation process itself (de Solla Price, 1976) and to identify the factors that affect the citedness of academic papers (Bornmann, Schier, Marx, & Daniel, 2012; Didegah & Thelwall, 2013). For example, negative binomial regression previously has been used to analyse factors underlying patent citations (Maurseth & Verspagen, 2002). The choice of statistical model is not straightforward (Bookstein, 2001), however, because citation data is typically highly skewed (de Solla Price, 1976) with a heavy tail (i.e., with particularly many articles having high citation counts) which makes it difficult to identify and fit the best distribution (Clauset, Shalizi, & Newman, 2009). Nevertheless, it has recently been shown that the distribution of citations to articles from an individual Scopus category and year follows a hooked power law or a discretised lognormal distribution substantially better than a power law (Thelwall & Wilson, 2014a) and that, on this basis, (discretised) ordinary least squares regression on the log of the citation data, after adding 1 to cope with the problem of uncited articles, is applicable and is probably the best available regression method (Thelwall & Wilson, 2014b). It should be noted that although the data is well fitted by the discretised lognormal distribution, it should not be assumed that it was derived from that distribution, as models should not be regarded as literal descriptions of nature (Hesse, 1953). Moreover, it is useful to assess additional statistical models in case a more powerful model can be found as well as to shed light on the processes underlying citation, which are still far from fully understood. This paper investigates stopped sum models for citation data for the first time. These have very different underlying assumptions to the lognormal distribution but can result in similar shaped distributions.

Hence, should citation data fit them well, the results would have both practical and theoretical implications for citation analysis.

Stopped sum distributions

Stopped sum distributions were initially developed by Neyman to model the number of larvae in a field (Neyman, 1939). Neyman viewed the distribution of larvae as resulting from two population waves. The first 'parent' (or primary wave) distribution was followed by a distribution of 'offspring' (or secondary wave), whereby the numbers in the secondary wave would be dependent on the numbers in the primary wave; the overall population being the sum of the populations from the two waves (Johnson, Kemp, & Kotz, 2005, pp. 381–382). The two waves can have completely different statistical distributions. If, for example, the primary wave distribution is Poisson and the secondary wave distribution is negative binomial, the overall distribution is known as a *Poisson stopped sum negative binomial (NB) distribution*. Here stopped sum models are explored due to their potential to model citation data as two waves, the primary wave and secondary wave. Given that the overall number of citations that an article receives might come from a similar two waves process, the primary wave representing citations received shortly after a journal article has been published, and the secondary wave, perhaps overlapping with the first to some extent, representing the citations received as a result of scientists discovering an article because of its previous citations, either directly by following citations or indirectly because more cited articles are ranked more highly in some citation databases.

The stopped sum models for citation counts could also be appropriate if the two waves occurred simultaneously instead of sequentially. For example, for the Poisson stopped sum negative binomial model, one of the wave distributions follows the Poisson distribution and the other wave follows the negative binomial distribution at the same time.

The original model proposed by Neyman (1939) assumed that zero counts in the primary wave will automatically be followed by zero counts in the second wave. Hence, if X follows the Poisson stopped sum NB distribution, P(X=0) is just P(X=0) under the Poisson distribution.

For citation counts of one or more, the stopped sum assumes that this can only be a result of a non-zero citation in the primary wave. For example, a citation count of 3 can only arise as a result of one of the three combinations:

- 3 citations in the primary wave, 0 citation in the secondary wave; or
- 2 citations in the primary wave, 1 citation in the secondary wave; or
- 1 citation in the primary wave, 2 citations in the secondary wave.

The Poisson stopped sum NB distribution will therefore have the following probability mass function (p.m.f.):

$$P(X=y)=\begin{cases} e^{-\lambda} & if\ y=0 \\ \sum_{j=1}^{y}\frac{e^{-\lambda}\lambda^{j}}{j!} * \binom{y-j+\alpha-1}{\alpha-1} p^{\alpha}(1-p)^{y-j} & if\ y\geq 1, \text{and } p=\frac{\alpha}{\mu+\alpha} \end{cases}$$

The other stopped sum distributions that are considered include the NB stopped sum Poisson distribution:

$$P(X=y)=\begin{cases} p^{\alpha} & if\ y=0 \\ \sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\frac{e^{-\lambda}\lambda^{y-j}}{(y-j)!} & if\ y\geq 1 \end{cases}$$

and the NB stopped sum NB distribution:

$$P(X=y)=\begin{cases} p^{\alpha} & if\ y=0 \\ \sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\binom{y-j+\theta-1}{\theta-1}q^{\theta}(1-q)^{y-j} & if\ y\geq 1 \end{cases}$$

where $p=\frac{\alpha}{\mu+\alpha}$ in all cases.

The Poisson stopped sum Poisson distribution was considered but because very large AICs were obtained indicating a poor fit for citation data we do not discuss it further here.

Modified stopped sum distributions

In the study made by Neyman in 1939, the restriction of having zero counts in the primary wave resulting in zero counts in the secondary wave was necessary, but in the case of citation analysis, it is feasible that a zero citation count in the first population wave could be followed by a non-zero count in the second. This can occur due to the limitations of the citation database used to analyse the citations. For example, an article may be uncited in Scopus, but cited in Google Scholar, and its Google Scholar citations could attract new second wave citations. Hence a modified stopped sum is also considered, where, for example, 3 citations could arise from 0 citations in the primary wave and 3 citations in the secondary wave. The modified Poisson stopped sum NB distribution for this case has p.m.f.:

$$P(X=y)=\sum_{j=0}^{y}\frac{e^{-\lambda}\lambda^{j}}{j!}*\binom{y-j+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y-j} \quad where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Using similar adjustments, the modified NB stopped sum Poisson distribution has p.m.f.:

$$P(X=y)=\sum_{j=0}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\frac{e^{-\lambda}\lambda^{y-j}}{(y-j)!} \quad where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Whilst the modified NB stopped sum NB distribution has p.m.f.:

$$P(X=y)=\sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\binom{y-j+\theta-1}{\theta-1}q^{\theta}(1-q)^{y-j}$$

$$where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Note that the modified Poisson stopped sum Poisson distribution is equivalent to a Poisson distribution, and hence is not considered here.

Research Questions

1. Do stopped sum models fit citation count data better than discretised lognormal and negative binomial models?
2. If so, which stopped sum model produces the most consistent results?

Methods

Data from 20 different subject areas were selected from Scopus in order to assess the models for a wide range of different disciplines. This is important because citation patterns are known to vary considerably between disciplines. This data has previously been analysed in Thelwall and Wilson (2014). Each subject area is a single Scopus category and consists of all documents of type article that were published in 2004, giving ten years for the articles to attract citations.

Fitting statistical models

The models were fitted using the R software (R Core Team, 2014). The MASS package (Venables & Ripley, 2002) was used to fit the negative binomial distribution. As there are no known statistical packages readily available to model the proposed stopped sum distributions, the parameters of the distributions were estimated by maximum likelihood estimations methods. AIC is a commonly used statistic for model selection, the model with the lowest AIC usually being regarded as the model that best fits the data (Bozdogan, 2000).

$$AIC = -2 \times \log(L) + (2 \times p)$$

Hence the AIC may be regarded as a penalised version of the loglikelihood, where L is the likelihood of the model and p is the number of parameters estimated. For example, both the Poisson stopped sum NB and NB stopped sum Poisson will have p=3, as there is one parameter (λ) in the Poisson wave and two parameters (NB mean, μ and size, α) in the NB wave. The NB stopped sum NB model will have p=4 as two parameters (μ and α) are estimated in each of the NB waves. Whilst opinions differ, when selecting the 'best' model, it has been suggested that a difference of 6 between the AICs will be large enough to imply a significant difference between the models (Burnham & Anderson, 2003).

Standard errors

Standard errors were computed to reflect the precision with which the proposed statistical models estimate the relevant parameters (Dodge, 2003, p. 386). For the negative binomial models, standard errors were obtained directly from the model fitting software. For the discretised lognormal, the standard errors were obtained by bootstrapping.

For other models the standard errors were calculated using the Hessian matrix, which is the matrix of the second derivatives of the log-likelihood function. The Hessian matrix can also be obtained whilst estimating the parameters for the corresponding distributions using the optim function in R (R Core Team, 2014). Suppose that L represents the log-likelihood function of a stopped sum distribution with two parameters, say λ and μ, then the Hessian matrix is given by $\begin{pmatrix} \frac{\partial^2 L}{\partial \lambda^2} & \frac{\partial^2 L}{\partial \mu \partial \lambda} \\ \frac{\partial^2 L}{\partial \mu \partial \lambda} & \frac{\partial^2 L}{\partial \mu^2} \end{pmatrix}$, and the standard errors for λ and μ are calculated as the square root of the main diagonal of the inverse of the negative Hessian matrix (Ruppert, 2011, pp. 166–167). At 95% confidence interval can be computed by parameter estimate ± 1.96*standard error.

Results

The modified negative binomial stopped sum negative binomial distribution (NBNB) produced the lowest AIC for 13 out of 20 subjects. The next most successful models are the NB stopped sum NB and the discretised lognormal. The Poisson stopped sum NB and the modified NB stopped sum Poisson each fitted 'best' for only one subject (see Table 3 in Appendix).

Parameter estimates for stopped sum distributions

The estimated parameters for Tourism and Soil will be discussed for the proposed stopped sum distributions. These subjects were selected as they are examples of subjects, which return parameter estimates and errors for all the fitted distributions. From Table 1, when Tourism is fitted with the Poisson stopped sum NB model, one wave follows the Poisson distribution with mean, λ=3.22, whilst the other wave follows a negative binomial distribution with mean, μ=18.77 and size, α=0.57; thus the negative binomial wave has a variance of 640.19, since the negative binomial variance equals $\frac{\mu^2}{\alpha} + \mu$. However, when fitted with the NB stopped sum Poisson model, one wave follows a negative binomial distribution with mean, μ=21.53, size, α=0.98, and variance=495.77, whilst the other wave follows a Poisson distribution with mean, λ=0.01. The estimated means (μ) in both negative binomial waves are relatively larger than the estimated means (λ) in the Poisson waves, suggesting that the majority of citation counts for Tourism derive from the negative binomial wave. This supports the interpretation that the two waves occur simultaneously, instead of sequentially, as mentioned above. It is also interesting to note that the sum of the estimated means from the Poisson waves and negative binomial waves of these stopped sum models are approximately equal to the estimated mean when Tourism is fitted solely with the negative binomial model.

When fitted with the NB stopped sum NB model, the estimated mean for Tourism in the primary NB wave (13.48) is larger than that of the secondary NB wave (8.25), suggesting that the majority of citation counts for Tourism derive from the primary wave. Furthermore, the sum of the estimated means from the NB stopped sum NB model for Tourism is also approximately equal to the estimated mean when Tourism is fitted with the negative binomial model only.

Similar results were obtained for Soil. When citation counts for Soil are fitted with the Poisson stopped sum NB model and NB stopped sum Poisson model, the mean estimates in the NB waves are much larger than those of the Poisson waves, suggesting that the majority of citation counts from Soil derive from the NB wave. Moreover, the sum of the estimated means for the stopped sum models is approximately equal to the estimated mean for the negative binomial model only (which is 16.93).

Table 1. Estimated parameters for the NB, Poisson stopped sum NB, NB stopped sum Poisson and NB stopped sum NB models.

	Negative binomial		*Poisson stopped sum NB*			*NB stopped sum Poisson*			*NB stopped sum NB*			
Sub.	*mu*	*size*	λ_1	*mu2*	*size2*	*mu1*	*size1*	λ_2	*mu1*	*size1*	*mu2*	*size2*
Tour.	21.53	0.98	3.22	18.77	0.57	21.53	0.98	0.01	13.48	1.30	8.25	0.10
Soil	16.93	0.74	2.27	16.09	0.56	16.87	0.74	0.06	13.78	0.82	3.46	0.04

Table 2 compares estimated parameters for the NB distribution against those of the modified stopped sum distributions. For the modified versions, the estimates of the Poisson stopped sum NB are similar to those of the NB stopped sum Poisson distributions. Similarly to the stopped sum distributions, Tourism and Soil depends largely on the wave that derives from

the NB distribution, as the λ estimates are relatively lower than the mu estimates. Furthermore, the sum of the two mu estimates for the modified NB stopped sum NB distributions (21.533 and 16.931) are also similar to the estimates from the NB distribution.

Table 2. Estimated parameters for the NB, modified Poisson stopped sum NB, modified NB stopped sum Poisson and modified NB stopped sum NB models.

	Negative binomial		*Modified Poisson stopped sum NB*			*Modified NB stopped sum Poisson*			*Modified NB stopped sum NB*			
Subj.	*mu*	*size*	λ_1	*mu2*	*size2*	*mu1*	*size1*	λ_2	*mu1*	*size1*	*mu2*	*size2*
Tour.	21.53	0.98	1.41	20.12	0.75	20.12	0.75	1.41	14.75	0.35	6.79	1.17
Soil	16.93	0.74	0.11	16.82	0.72	16.81	0.72	0.11	4.92	0.08	12.01	0.75

Standard errors for stopped sum distributions

Figures 1 and 2 show the mean and size estimates for the primary and secondary waves of the modified NB stopped sum NB distributions. Visual, Literature and Rehab were excluded as standard errors could not be obtained as a result of a singular hessian matrix.

Although the modified NB stopped sum NB distribution gave the lowest AIC, the model produced very large standard errors, resulting in large confidence intervals, as shown in Figures 1 and 2, indicating that this modified NB stopped sum NB model is impractical. This result could possibly be due to the nature of citations, which differs from that of the larvae studied by Neyman. With larvae and their offspring it is clear which wave of population a larvae originates from, this is not the case with citations – usually it will be far from clear cut which wave a given citation might belong to, which in turn leads to difficulty estimating the mean number of citations for that wave, and hence the large associated standard errors.

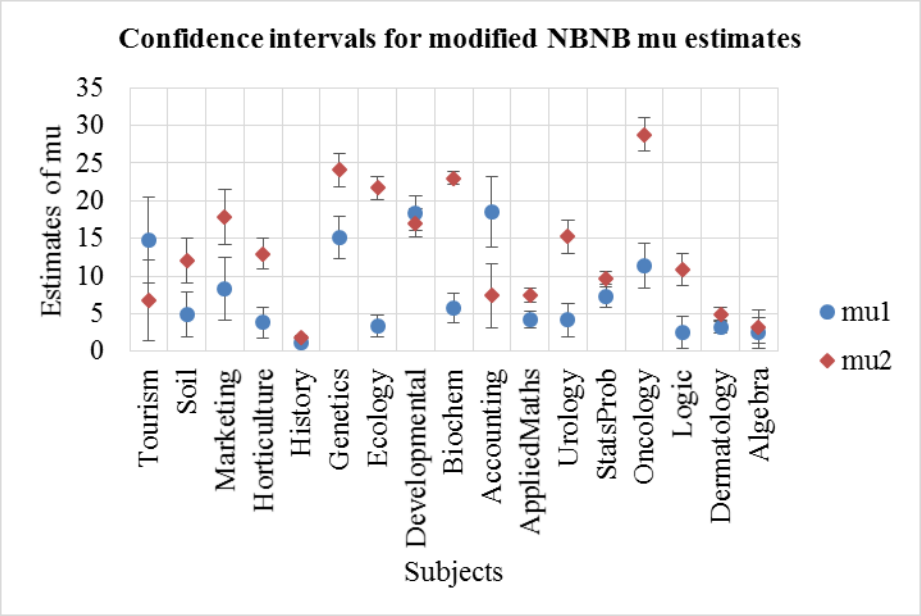

Figure 1. Mean (mu) estimates for the modified NB stopped sum NB distribution for both primary and secondary waves with 95% confidence intervals.

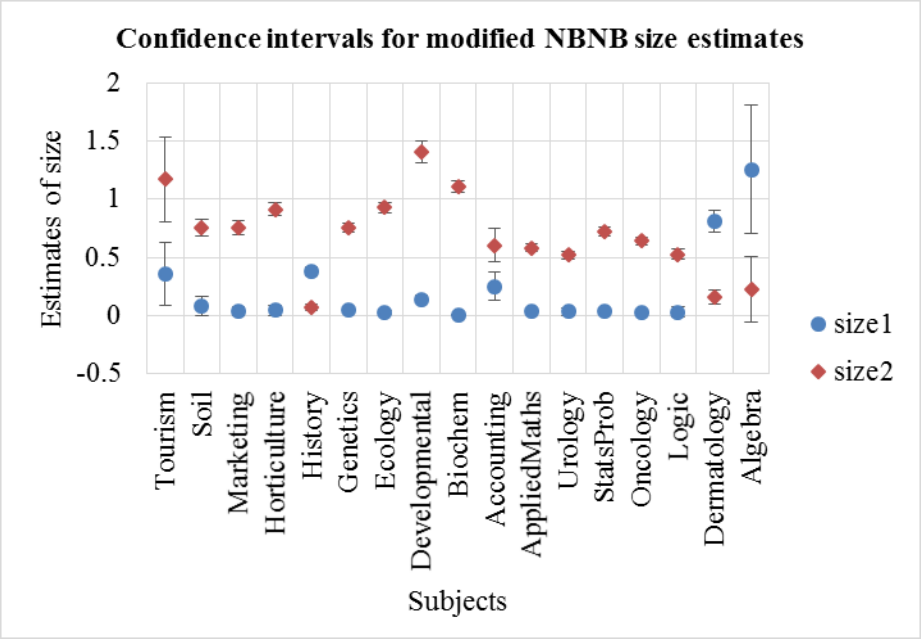

Figure 2. Size estimates for the modified NB stopped sum NB distribution for both primary and secondary waves with 95% confidence intervals.

A further examination of the modified NBNB stopped sum model was carried out with simulations using some known fixed parameters, and similar results were obtained. Moreover, simulations were carried out on all the other stopped sum models and similar results were also obtained for the NBNB stopped sum distribution. Hence it can be concluded that both the stopped sum and modified NBNB stopped sum models are impractical when modelling data with no covariates. Further studies should be conducted to see if adding covariates would change the reliability of the model.

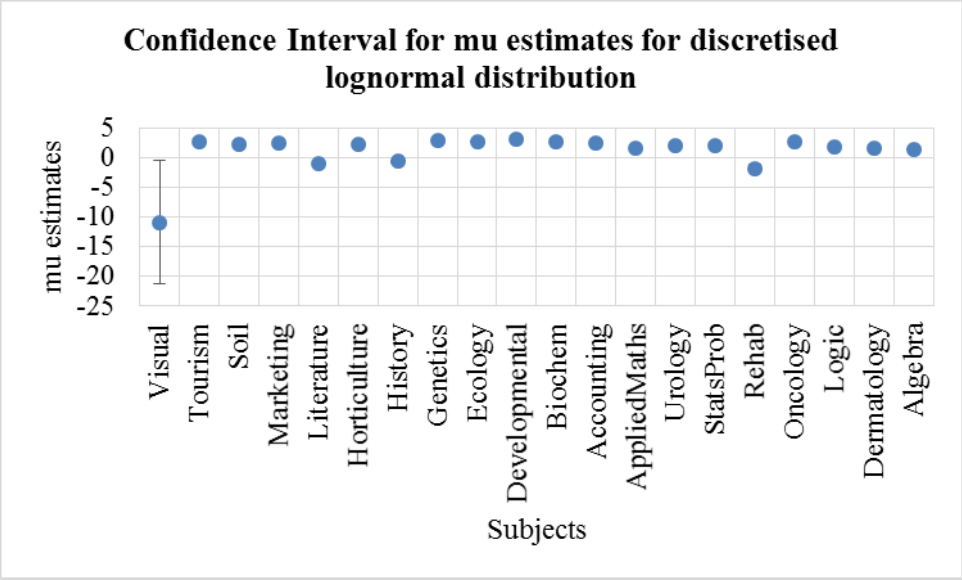

Figure 3. Mu estimates for the discretised lognormal distribution with 95% confidence intervals.

On the other hand, the 95% confidence interval for all subjects except Visual for the discretised lognormal distribution (Fig. 3) are much narrower compared to that of the

modified NB stopped sum NB distribution. This indicates that the discretised lognormal distribution is more suitable in practice.

Conclusions

This paper tested stopped sum distributions for modelling citation data for the first time and also introduces a modification to allow the 'waves' to occur simultaneously rather than sequentially. However, given that the standard errors for the stopped sum distribution tend to be very large it is doubtful whether these distributions are useful for citation data even though they produce the lowest AIC. For example, out of all the tested distributions, the modified NB stopped sum NB distribution produced the lowest AIC, but the large standard errors suggests that it is an unsuitable model as its parameter estimates are too unreliable for predictions or conclusions based upon the model to be meaningful.

Overall, the results suggest that for covariate free data, the discretised lognormal distribution is much more suitable for regressing citation data from a single subject and year. Nevertheless, on a theoretical level, the good fits found for some of the stopped sum models give evidence that there are (at least) two important and separate processes that govern the citing practices of authors. For one of these processes, existing citations are irrelevant for new citations, and for the other, they are relevant.

References

Bookstein, A. (2001). Implications of ambiguity for scientometric measurement. *Journal of the American Society for Information Science and Technology*, *52*(1), 74–79. doi:10.1002/1532-2890(2000)52:1<74::AID-ASI1052>3.0.CO;2-C

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, *6*(1), 11–18.

Bozdogan, H. (2000). Akaike's Information Criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, *44*(1), 62–91. doi:10.1006/jmps.1999.1277

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd ed., p. 520). Springer.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. doi:10.1137/070710111

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. doi:10.1002/asi.4630270505

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, *7*(4), 861–873. doi:10.1016/j.joi.2013.08.006

Dodge, Y. (2003). *The Oxford dictionary of statistical terms*. (S. D. Cox, D. Commenges, A. Davison, P. Solomon, & S. Wilson, Eds.) (1st ed., p. 498). Oxford: Oxford University Press.

Hesse, M. B. (1953). Models in Physics. *The British J. for the Philosophy of Science*, *4*(15), 198–214.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distribution* (3rd ed., p. 672). Wiley-Interscience.

Maurseth, P. B., & Verspagen, B. (2002). Knowledge spillovers in Europe: A patent citations analysis. *Scandinavian Journal of Economics*, *104*(4), 531–545. doi:10.1111/1467-9442.00300

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, *10*(1), 35–57. doi:10.1214/aoms/1177732245

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Ruppert, D. (2011). *Statistics and data analysis for financial engineering* (p. 638). New York: Springer.

Thelwall, M., & Wilson, P. (2014a). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839. doi:10.1016/j.joi.2014.08.001

Thelwall, M., & Wilson, P. (2014b). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971. doi:10.1016/j.joi.2014.09.011

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Appendix

Table 3. AIC for all subjects for each distribution

Subjects	*Discretised lognormal*	*Negative binomial*	*Poisson stopped sum NB*	*NB stopped sum Poisson*	*NB stopped sum NB*	*Modified Poisson stopped sum NB*	*Modified NB stopped sum Poisson*	*Modified NB stopped sum NB*	*Number of articles*
Visual	7902	7928	7916	7930	7865	7920	7920	**7865**	4096
Tourism	4956	4980	4980	4982	4969	4964	4964	**4955**	608
Soil	33470	33344	33458	33345	33287	33344	33344	**33282**	4347
Marketing	**12917**	13073	13025	13073	12941	13015	13015	12932	1550
Literature	11624	11635	**11618**	11637	11622	104485	11624	25449	5000
Horticulture	23058	23093	23165	23095	23001	23067	23067	**22992**	3009
History	19797	19994	19849	19996	19824	19880	19880	**19795**	5000
Genetics	45622	46014	45997	46002	45474	45982	45982	**45471**	5000
Ecology	42787	42343	42441	42335	42253	42366	42793	**42240**	5000
Developmental	40985	41604	41340	41558	40979	41385	41385	**40956**	4541
Biochem	42901	43690	43540	43638	**42675**	43659	43659	42680	5000
Accounting	9927	9933	9924	9931	9914	9929	9929	**9896**	1178
AppliedMaths	33504	33739	33704	33741	33460	33685	33685	**33441**	5000
Urology	38932	38621	38793	38623	**38560**	38623	38623	38563	5000
StatsProb	**36696**	37416	37177	37418	36742	37186	37186	36706	5000
Rehab	28086	27531	27622	27533	27628	27483	**27483**	28322	5000
Oncology	42577	42620	42679	42607	**42196**	42660	42684	42225	4646
Logic	32258	32044	32164	32046	32012	32045	32045	**32010**	4547
Dermatology	19608	19774	19671	19776	19675	19692	19692	**19606**	3184
Algebra	**2968**	2991	2973	2993	2977	2978	2978	2972	528

Table 4. Estimated parameters of negative binomial distribution with the stopped sum distributions

	Negative binomial		*Poisson stopped sum NB*			*NB stopped sum Poisson*			*NB stopped sum NB*			
Subjects	*mu*	*size*	*lambda1*	*mu2*	*size2*	*mu1*	*size1*	*lambda2*	*mu1*	*size1*	*mu2*	*size2*
Visual	0.66	0.17	0.28	1.61	0.34	0.66	0.17	0.00	0.60	0.19	0.26	0.00
Tourism	21.53	0.98	3.22	18.77	0.57	21.53	0.98	0.01	13.48	1.30	8.25	0.10
Soil	16.93	0.74	2.27	16.09	0.56	16.87	0.74	0.06	13.78	0.82	3.46	0.04
Marketing	26.13	0.63	2.63	24.97	0.43	26.02	0.62	0.12	20.34	0.76	6.16	0.01
Literature	0.79	0.32	0.40	1.18	0.33	0.79	0.32	0.00	0.41	9.22	1.16	0.31
Horticulture	16.72	0.83	2.52	15.15	0.54	16.71	0.83	0.01	14.27	0.94	2.62	0.02
History	2.90	0.30	0.75	4.08	0.27	2.90	0.30	0.00	1.26	0.75	3.12	0.12
Genetics	39.23	0.61	2.71	38.78	0.50	38.96	0.60	0.28	24.30	0.80	15.85	0.04
Ecology	25.02	0.86	2.52	24.17	0.79	24.73	0.84	0.31	22.61	0.76	2.60	0.32
Developmental	35.45	0.93	4.03	31.86	0.60	34.56	0.86	0.90	17.95	1.52	17.73	0.12
Biochem	28.81	0.84	3.21	26.60	0.61	28.08	0.79	0.75	22.86	1.12	6.09	0.01
Accounting	25.89	0.64	2.46	25.36	0.50	25.66	0.63	0.26	12.93	0.87	14.03	0.12
AppliedMaths	11.71	0.50	1.68	12.20	0.39	11.71	0.50	0.00	8.20	0.63	4.28	0.03
Urology	19.39	0.51	1.80	20.69	0.50	19.47	0.51	0.00	15.49	0.56	4.60	0.03
StatsProb	16.93	0.54	2.12	16.62	0.36	16.93	0.54	0.00	10.50	0.77	7.21	0.03
Rehab	9.29	0.23	0.83	14.56	0.37	9.28	0.23	0.00	0.83	89.55	14.56	0.37
Oncology	40.23	0.55	2.34	41.68	0.53	39.94	0.54	0.33	25.50	0.68	16.33	0.05
Logic	13.40	0.53	1.67	14.21	0.49	13.37	0.53	0.00	11.59	0.56	2.19	0.02
Dermatology	8.07	0.65	1.79	7.44	0.37	8.06	0.65	0.01	1.83	41.25	7.39	0.36
Algebra	5.75	0.90	1.90	4.46	0.37	5.74	0.90	0.01	1.94	42.31	4.41	0.36

Table 5. Estimated parameters of negative binomial distribution with the modified stopped sum distributions

	Negative binomial		*Modified Poisson stopped sum NB*			*Modified NB stopped sum Poisson*			*Modified NB stopped sum NB*			
Subjects	*mu*	*size*	*lambda1*	*mu2*	*size2*	*mu1*	*size1*	*lambda2*	*mu1*	*size1*	*mu2*	*size2*
Visual	0.66	0.17	0.04	0.62	0.14	0.62	0.14	0.04	0.60	0.19	0.06	0.00
Tourism	21.53	0.98	1.41	20.12	0.75	20.12	0.75	1.41	14.75	0.35	6.79	1.17
Soil	16.93	0.74	0.11	16.82	0.72	16.81	0.72	0.11	4.92	0.08	12.01	0.75
Marketing	26.13	0.63	1.02	25.11	0.50	25.11	0.50	1.02	8.35	0.03	17.78	0.76
Literature	0.79	0.32	11.82	11.99	0.00	0.72	0.24	0.07	4.65	2.71	3.85	0.00
Horticulture	16.72	0.83	0.50	16.24	0.73	16.18	0.72	0.53	3.82	0.05	12.90	0.91
History	2.90	0.30	0.20	2.70	0.21	2.70	0.21	0.20	1.08	0.38	1.82	0.07
Genetics	39.23	0.61	0.43	38.81	0.57	38.81	0.57	0.43	15.12	0.04	24.12	0.75
Ecology	25.02	0.86	0.00	23.60	0.91	18.21	0.80	0.00	3.36	0.02	21.67	0.93
Developmental	35.45	0.93	2.56	32.89	0.69	32.89	0.69	2.56	18.40	0.14	17.04	1.41
Biochem	28.81	0.84	0.69	28.12	0.76	28.12	0.76	0.69	5.79	0.01	23.02	1.11
Accounting	25.89	0.64	0.34	25.55	0.60	25.55	0.60	0.34	18.48	0.25	7.40	0.60
AppliedMaths	11.71	0.50	0.28	11.44	0.44	11.44	0.44	0.28	4.26	0.04	7.45	0.58
Urology	19.39	0.51	0.02	19.37	0.51	19.37	0.51	0.02	4.17	0.03	15.21	0.52
StatsProb	16.93	0.54	0.78	16.16	0.41	16.15	0.41	0.78	7.19	0.04	9.74	0.72
Rehab	9.29	0.23	0.09	9.19	0.21	9.19	0.21	0.09	5.71	0.00	25.74	0.20
Oncology	40.23	0.55	0.00	45.66	0.54	34.70	0.57	0.00	11.43	0.02	28.81	0.64
Logic	13.40	0.53	0.04	13.37	0.52	13.37	0.52	0.04	2.52	0.03	10.88	0.53
Dermatology	8.07	0.65	0.60	7.48	0.47	7.48	0.47	0.60	3.22	0.81	4.85	0.16
Algebra	5.75	0.90	0.84	4.91	0.55	4.91	0.55	0.84	2.48	1.25	3.27	0.23