Scientometrics (2020) 125:993–1010
https://doi.org/10.1007/s11192-020-03654-8

A new algorithm for zero-modified models applied to citation counts

Marzieh Shahmandi[1] · Paul Wilson[1] · Mike Thelwall[1]

Received: 13 December 2019 / Published online: 17 August 2020

Abstract

Finding statistical models for citation count data is important for those seeking to understand the citing process or when using regression to identify factors that associate with citation rates. As sets of citation counts often include more or less zeros (uncited articles) than would be expected under the base distribution, it is essential to deal appropriately with them. This article proposes a new algorithm to fit zero-modified versions of discretised log-normal, hooked power-law and Weibull models to citation count data from 23 different Scopus categories from 2012. The new algorithm allows the standard errors of all parameter estimates to be calculated, and hence also confidence intervals and p-values. This algorithm can also estimate negative zero-modification parameters corresponding to zero-deflation (fewer uncited articles than expected). The results find no universal best model for the 23 categories. A given dataset may be zero-inflated relative to one model, but zero-deflated relative to another. We suggest circumstances in which one of the models under consideration may be the best fitting model.

Keywords Zero-modified models · Discretised log-normal distribution · Hooked power-law distribution · Weibull distribution

Introduction

It is important to identify models that fit citation distributions well for several reasons. A good model can be used to identify anomalous sets of articles that are not fitted well by the model, suggesting indexing or classification errors, can help with the design of effective impact indicators and confidence intervals, and is important when performing regression analyses to identify factors that influence citations. A common problem when fitting statistical models to citation data is that the number of uncited articles (0 s) differs from that expected by the best fitting model, perhaps due to citation indexing policies selecting the wrong balance of high and low impact journals. This problem might be remedied by fitting a zero-inflated or a zero-deflated (i.e., a zero-modified) distribution that allows the predicted number of zeros to more closely approximate the number of zeros in a dataset.

✉ Marzieh Shahmandi
m.shahmandihounejani@wlv.ac.uk

[1] Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton, UK

A previous study fitted zero-inflated versions of the discretised log-normal and hooked power law distributions to citation count data from 23 Scopus categories, finding that zero-inflation occurred in nearly all cases (Thelwall 2016). The zero-inflation was hypothesised to be a consequence of “inherently unciteable articles”, such as magazine articles. Zero-counts due to unciteability are an example of “perfect” or “structural” zeros: data that are constrained to be zeros due to some feature of the data generating process. In contrast, other zeros are referred to as non-perfect or count zeros. In this context, a non-perfect zero would be a paper that is citeable but has not been cited. In essence, zero-inflated models seek to estimate the proportion of perfect zeros present in data and fit a count distribution to the remaining data.

A less well-studied phenomenon is zero-deflation, where data is well-fitted by a given count distribution, but there are less zeros present in the data than would be expected under the distribution. Zero-deflation may arise for citation counts from the Web of Science (WoS), Scopus or any other citation database with selective inclusion criteria because uncited articles may be less likely to be indexed. WoS and Scopus have poorer coverage of non-English journals than of English journals so the absence of non-English journals may contribute to zero-deflation. This may be particularly relevant for fields containing nation-specific agricultural, legal, culture, or politics research.

Whilst previous scientometric studies have fitted zero inflated distributions to citation count data, none have fitted zero-deflated or zero-modified distributions. This paper introduces zero-modified versions of the hooked power law and discretized log-normal distributions previously shown to fit citation data well (Thelwall 2016), and also zero-modified version of the discrete Weibull distribution. The discrete Weibull distribution is capable of modelling highly skewed count data with more zeros and thus is a good candidate model for citation counts (Brzezinski 2015). Discrete Weibull distributions may be fitted to data using the R-Package DWreg (Vinciotti 2016). The pure power law distribution is not considered because it usually requires low cited articles to be ignored for fitting and therefore is not a credible citation distribution. This paper also introduces an algorithm that fits both negative and positive zero-modification parameters and determines the standard errors of the zero-modification (and other) parameters, which in turn enables the calculation of confidence intervals for these parameters, and the performance of statistical tests on them. The algorithms are tested on a sample set of citation data from 23 fields to assess the extent to which the new distributions fit citation count data. The circumstances in which one of the models under consideration may be the best fit are also discussed.

Distributions

Hooked power law

The hooked power law is a generalised version of the power law model (Pennock et al. 2002). The hooked power law has probability mass function:

$$f(x;B,\alpha) = \begin{cases} A(B+x)^{-\alpha} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where B and α are model parameters, and A is a constant chosen so that $\sum_{x=0}^{\infty} f(x;B,\alpha) = 1$.

Discretized log-normal

A (continuous) random variable is log-normally distributed if its logarithm is normally distributed. It has probability density function:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), \quad x > 0, \sigma > 0, \mu \in (-\infty, +\infty). \tag{1}$$

To discretise the distribution, (i.e., convert it into a form that models the situation where x is a positive integer), integrate $f(x;\mu,\sigma)$ over unit intervals about positive integer values of x, and divide by $K = \int_{0.5}^{\infty} f(x;\mu,\sigma)\mathrm{d}x$, where f is as at (1) above. Thus, the probability mass function of the discretised log-normal distribution is:

$$g(x;\mu,\sigma) = \begin{cases} \frac{1}{K}\int_{x-0.5}^{x+0.5} f(x;\mu,\sigma)\mathrm{d}x & x = 1,2,3,\ldots \\ 0 & \text{otherwise} \end{cases}$$

Discrete Weibull

The discrete Weibull distribution has probability mass function:

$$f(x;q,\beta) = \begin{cases} q^{x^\beta} - q^{(x+1)^\beta} & x = 0,1,2,3,\ldots \\ 0 & \text{otherwise} \end{cases}$$

where $0 < q < 1$ and $\beta > 0$.

Zero-modified models

A zero-modified model (see, for example, Dietz and Böhning 2000) has the probability mass function:

$$f(x;\omega,\Theta) = \begin{cases} \omega + (1-\omega)f^*(x;\Theta) & x = 0 \\ (1-\omega)f^*(x;\Theta) & x = 1,2,3,\ldots \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where Θ is a set of parameters and $f^*(x;\Theta)$ is a probability mass function. For negative ω the distribution is known as a zero-deflated distribution and for positive ω, it is known as a zero-inflated distribution. For $\omega = 0$ the model reduces to the non-modified model, f^*, if $\omega = 1$ the model is a "zero-model'', i.e. one where all data are zero. Zero-inflated models are used to model data that has excess zeros (more zero counts than expected under the model f^*). For example, if 100 data points/observations follow a Poisson distribution with parameter 1, we would expect to observe about $100 \times e^{-1} = 36.8$ zeros. Substantially more zeros would therefore indicate possible zero-inflation. Zeros may stem from two distinct processes, "Non-excess'' zeros where zeros occur by chance, in the same manner as 1 s, 2 s,…; and another process by which some data are constrained to be zeros (perfect or structural zeros).

For a zero-deflated model, $\omega < 0$, but may take values < -1. To see this, note that

$$\begin{aligned} &f(0;\omega,\Theta) \geq 0 \Leftrightarrow \omega + (1-\omega)f^*(0;\Theta) \geq 0 \\ &\Rightarrow \omega(1-f^*(0;\Theta)) + f^*(0;\Theta) \geq 0 \\ &\Rightarrow \omega > \tfrac{-f^*(0;\Theta)}{1-f^*(0;\Theta)} \end{aligned} \tag{3}$$

For example, if f^* is a Poisson distribution with parameter 0.5 then $f^*(0;0.5) = \exp(-0.5) = 0.6065$ and hence ω is valid provided

$$\omega \geq -\frac{0.6065}{1-0.6065} = -1.54$$

The interpretation of negative values of ω is not as straightforward as those of positive values. The most straightforward interpretation is to regard $1-\omega$ as the proportionate increase in the expected number of observed positive values. For example, if $\omega = -1.5$, then we would expect to observe approximately $1-(-1.5) = 2.5$ times more 1 s, 2 s, 3 s etc. in the data than we would in the non-modified model. Zero-deflation in data can be a consequence of some zero-counts not being included (e.g., Mendonca 1995).

Data and methods

The data analysed in this article consist of citation counts for journal articles published in 2012 from 23 Scopus categories, with up to 5000 journal articles for most of the categories. The citation counts were downloaded from Scopus in November 2017. If there were greater than 5000 articles in a category the most recent 5000 articles were selected. This provides a coherent collection of articles with 5–6 years of citations (see "Appendix 1").

A previously published algorithm fits zero-inflated discrete log-normal and zero-inflated hooked power law models to covariate-free data (the zero-inflation parameter is estimated to two decimal places) (Thelwall 2016). This model is easily extended to zero-inflated versions of any count model but is unable to fit negative zero-modification parameters. In this article, we propose an algorithm that will enable the fitting of negative (and positive) values of ω, will estimate the value of ω to many decimal places, and is much faster. R code to fit the models discussed in this paper is available online.[1]

This algorithm is based upon maximization of the log-likelihood of the relevant zero-modified models via the *optim* command of R. The *optim* function offers different optimisation algorithms, including conjugate gradient, quasi-Newton, Nelder–Mead and simulated annealing. The default method is a derivative-free Nelder–Mead algorithm that does not require the computation of the gradient. It also is a method for solving high-dimensional linear optimisation problems with constraints that is non-sensitive and robust to discontinuities in the likelihood surface, and generally requires relatively few function evaluations to achieve convergence.

The above-mentioned algorithms have several advantages over techniques such as Newton–Raphson and Fisher Scoring. In particular, they optimise the log-likelihood function

[1] The R source code is available at https://doi.org/10.6084/m9.figshare.7643093.v1.

of the parameters simultaneously as opposed to individually. Consequently, the estimators obtained are better than those obtained by maximising the likelihood with respect to each parameter. Such techniques have been around a long time but have only become practical in recent years due to improved computing power.

An advantage of the *optim* command is its output. It includes the parameter estimates (including the estimate of the zero-modification parameter), the value of the likelihood of the model, and the Hessian matrix (Faraway 2005). The diagonal entries of the matrix inverse are proportional to the standard errors of the parameter estimates. The issue of whether there is zero inflation or deflation is dealt with by the sign of the estimate of the zero-modification parameter. The log-likelihood of the model is also outputted, which enables the calculation of corresponding Akaike Information Criterion (AIC) introduced by Akaike (1974), and thus enables the model fits to be compared. The model usually considered "best" is the one with the lowest AIC. The AIC is essentially an adjusted version of the log-likelihood, with the adjustment being to account for differing numbers of variables between models. The models corresponding to the discrete lognormal, hooked power law, and Weibull have to be fitted independently and the comparison carried out manually. The code supplied is easily modified for other distributions.

As is mentioned above, the standard errors of the parameter estimates can be computed from the Hessian matrix. This is especially useful for the calculation of confidence intervals for the zero-modification parameter (as well as any other parameters). For the computation of the confidence intervals related to the zero-modification parameter ω the formula $\hat{\omega} \pm \mathcal{Z}_{1-\frac{\alpha}{2}} * Se(\hat{\omega})$ is used where Se is the standard error of the maximum likelihood estimate of ω and $\mathcal{Z}_{1-\frac{\alpha}{2}}$ is the $\left[\left(1-\frac{\alpha}{2}\right) \times 100\right]$-th percentile of a standard normal distribution. Thus, for $\alpha = 0.05, \mathcal{Z}_{1-\frac{\alpha}{2}}=1.96$, values of ω between the interval's limits are compatible with the data.

Having the estimates of the standard deviations of the parameters also enables the performance of hypothesis tests related to the parameters. In particular it enables the test $H_0 : \omega = 0$ to determine whether there is statistical evidence of zero-modification in the data. Whilst American Statistical Association (Wasserstein and Lazar 2016) guidelines concerning the misuse of p-values and confidence intervals has led to debate about their use, the guidelines are primarily concerned with the misuse use of p-values and confidence intervals and do not advise their abandonment. Indeed, the guidelines state that "P-values can indicate how incompatible the data are with a specified statistical model''.

Several tests exist to test for zero-modification, including likelihood ratio tests, score tests, and the Wilson–Einbeck test (Wilson and Einbeck 2019). Note that whilst the Vuong test for non-nested models has been used as a test of zero-inflation, this is erroneous (Wilson 2015). This paper uses the Wald test (Wasserman 2006) to test: $H_0 : \omega = 0$ against the alternative: $H_1 : \omega \neq 0$ with $W = \frac{\hat{\omega}}{Se(\hat{\omega})}$. We employ the Wald test as it directly tests the significance of the estimate of the zero-modification parameter without necessitating the fitting of the non-zero-modified model.

Finally, as it was mentioned for assessment of the fitted model, the AIC is used to show whether one model fits the data set better than another when the models in question contain differing numbers of parameters or predictor. Because there isn't any information about the distribution of the AIC values, the non-parametric bootstrap method is used to compute confidence intervals for the AIC values of the mentioned models for all 23 categories, using the R package "Boot'' (Canty and Ripley 2019). The bootstrap method (Efron 1979) is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. For example, in our case for one subject, a sample with size n from the

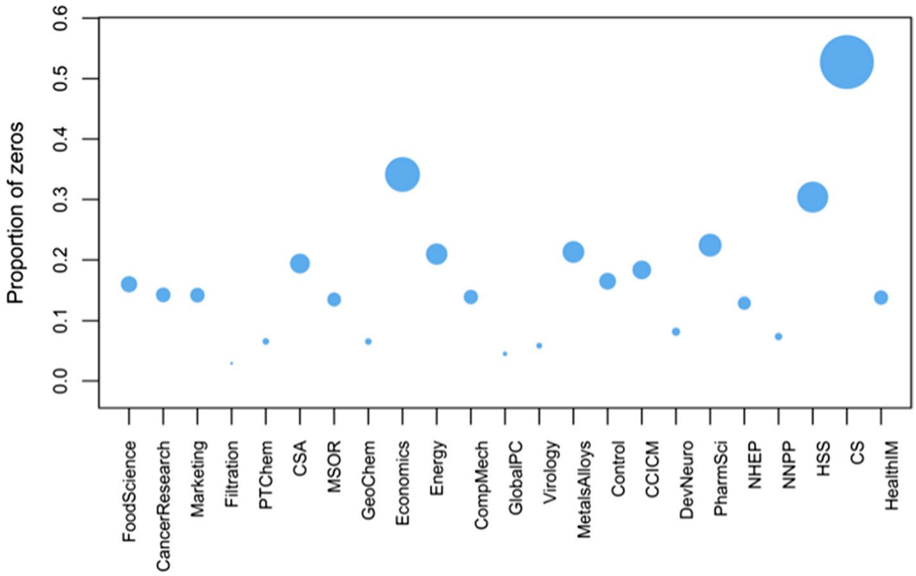

Fig. 1 The proportions of uncited articles (zeros) in citation data from 23 Scopus categories. Circle areas are proportional to proportions of zeros

data was drawn with replacement, and this was replicated B times. Each re-sampled sample of the data is considered as a bootstrap sample, so there are B bootstrap samples. For each bootstrap sample, the model is fitted, and AIC value is computed. There are B values of AIC and choosing the 2.5% and 97.5% percentiles gives a 95% percent confidence interval for AIC related to each zero-modified versions of the count distributions. This uses the R package "Boot''.

Results

Proportions of uncited articles

Uncited articles are far more common is some disciplines than in others (Fig. 1). Cultural Studies, Economics & Econometrics, Health Social Science, and Pharmaceutical Science have the greatest proportions of zero counts. In subjects such as Pharmaceutical Science large numbers of uncited articles might arise from publications which are not fully peer reviewed that might be regarded as magazines rather than journals being included in the database.

Zero-modified discretised log-normal distribution

The zero-modification parameter estimates for the discretised log-normal distribution are all positive, the largest estimates being for Health Social Science and Economics and the smallest for Filtration & Separation and Global & Planetary Change (Fig. 2, see also "Appendix 2"). There is almost universal zero-inflation relative to the discretized

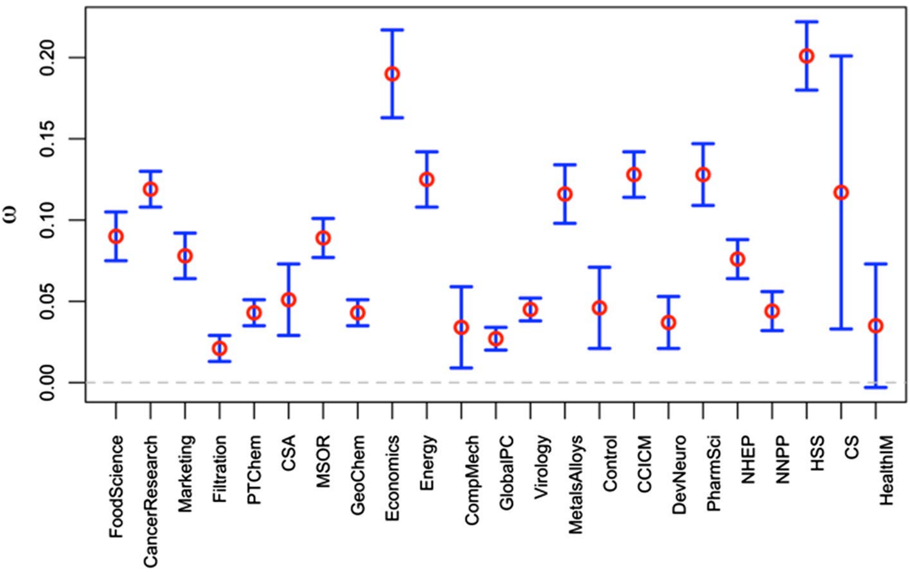

Fig. 2 Zero-modification parameters and 95% confidence intervals relative to a zero-modified discretised log-normal distribution for 23 Scopus categories

log-normal distribution. The zero-inflation parameter estimates for 22 of the 23 subjects are significant at a level of significance of $\alpha = 0.05$, with only Health Information Management returning a non-significant estimate: its confidence interval includes zero and its p value is also 0.07 which is larger than $\alpha = 0.05$ (it is clear that the confidence intervals and the p-values related to the parameter estimates are compatible, i.e. 0 is outside of the confidence interval when p = 0.05, and inside otherwise).

Zero-modified hooked power law distribution

Relative to a hooked power law distribution both significant positive (13 subjects) and significant negative (6 subjects) estimates of the zero-modification parameter occur, as well as 4 non-significant estimates (Fig. 3, see also "Appendix 3"). There is both zero-inflation and zero-deflation, and possibly no zero-modification relative to the hooked power law distribution.

Zero-modified discretised Weibull distribution

Relative to a discrete Weibull distribution only one estimate of the zero-modification parameter is significantly positive, 15 being significantly negative and 7 non-significant (Fig. 4, see also "Appendix 4"). There is both zero-inflation and zero-deflation and possibly no zero-modification relative to the discretise Weibull distribution.

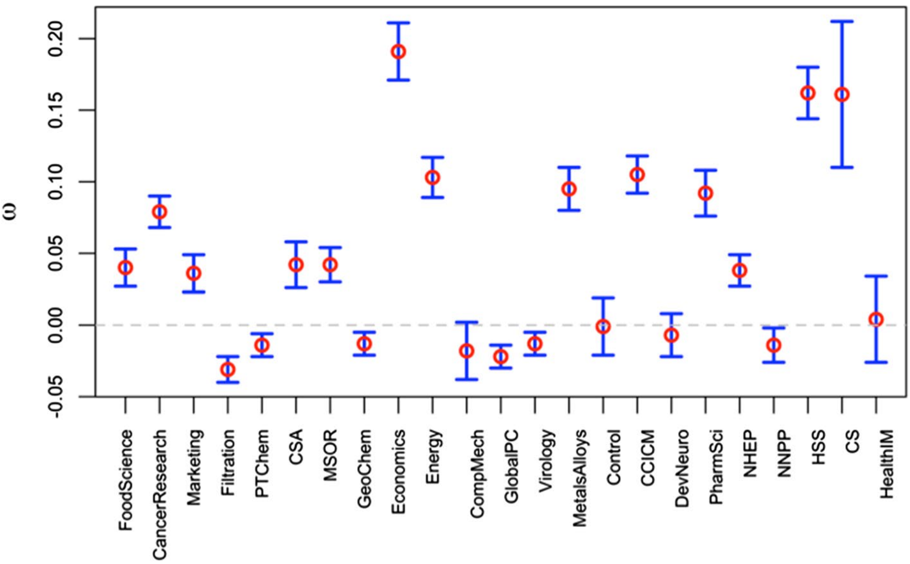

Fig. 3 Zero-modification parameters and their confidence intervals relative to a zero-modified hooked power law distribution for 23 Scopus categories

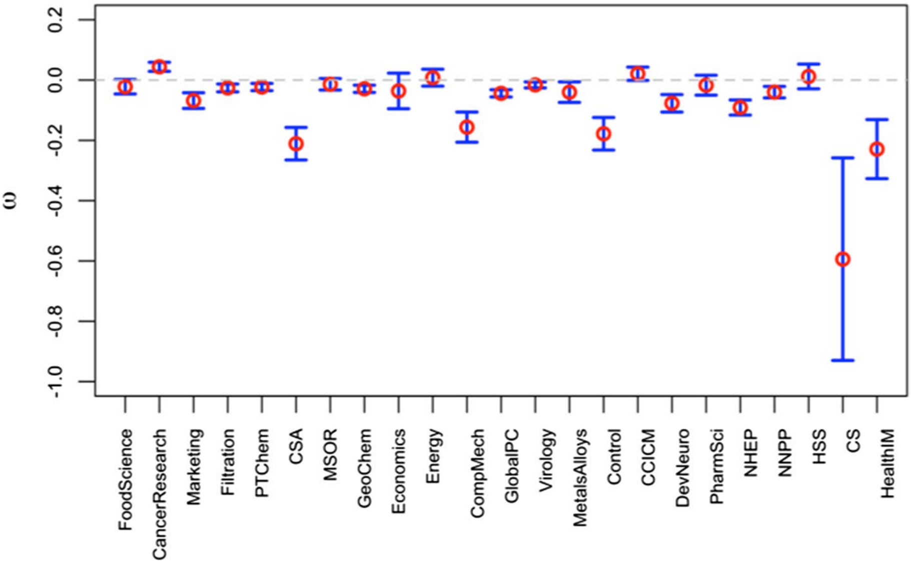

Fig. 4 Zero-modification parameters and their confidence intervals relative to a zero-modified discrete Weibull distribution for 23 Scopus categories

Table 1 AIC values and the estimates of zero-modification parameter for zero-modified version of discretised log-normal, hooked power law and Weibull for 23 Scopus categories

Subjects	AIC ZMDLN	ω	AIC ZMHPL	ω	AIC ZMWeibull	ω
Food science	31,450.58	0.09	**31,395.14**	0.04	31,400.72	−0.02
Cancer research	37,463.12	0.12	**37,427.02**	0.08	37,536.54	0.04
Marketing	32,368.34	0.08	**32,350.92**	0.04	32,453.12	−0.07
Physical and theoretical chemistry	36,178.02	0.04	**36,157.1**	−0.01	36,202.76	−0.02
Management science and operations research	34,601.58	0.09	**34,578.70**	0.04	34,636.12	−0.01
GeoChemistry and petrology	36,669.80	0.04	**36,663.82**	−0.01	36,740.60	−0.03
Computational mechanics	15,566.98	0.03	**15,561.42**	−0.02	15,593.68	−0.16
Control and optimization	18,377.46	0.05	**18,371.88**	0.00	18,422.50	−0.18
Developmental neuroscience	14,552.92	0.04	**14,528.48**	−0.01	14,581.18	−0.08
Nuclear and high energy physics	35,598.62	0.08	**35,556.86**	0.04	35,787.94	−0.09
Neuropsychology and physiological psych	20,238.10	0.04	**20,221.0**	−0.01	20,268.78	−0.04
Health social science	27,140.34	0.20	**27,119.98**	0.16	27,185.22	0.01
Health information management	6774.30	0.04	**6770.46**	0.00	6813.52	−0.23
Economics and econometrics	26,955.04	0.19	26,972.90	0.19	**26,935.90**	−0.04
Energy engineering and power technology	31,843.10	0.13	31,815.28	0.10	**31,750.70**	0.01
Metals and alloys	31,546.04	0.12	31,528.70	0.10	**31,514.86**	−0.04
Critical care and intensive care medicine	35,469.04	0.13	35,434.32	0.11	**35,430.70**	0.02
Pharmaceutical science	29,747.64	0.13	29,723.02	0.09	**29,717.20**	−0.02
Cultural studies	16,747.46	0.12	16,752.88	0.16	**16,746.14**	−0.59
Filtration and separation	**13,471.5**	0.02	13,537.42	−0.03	13,555.32	−0.03
Computer science application	**31,556.04**	0.05	31,564.34	0.04	31,596.68	−0.21
Global and planetary change	**29,961.92**	0.03	29,988.90	−0.02	30,091.84	−0.04
Virology	**37,268.22**	0.05	37,371.72	−0.01	37,450.60	−0.02

Best fitting distributions are in bold

Assessment of the models based on AIC

None of the models under consideration fits best in all cases (Table 1), the zero-modified hooked power law being best (under the AIC criterion) in 13 cases, the zero-modified Weibull in 6 cases and the zero-modified discrete log-normal in 4 cases. Table 1 suggests the best of the models under consideration for a given subject area. For 21 out of the 23 categories the estimate of the zero-modification parameter for the model with the lowest AIC is in the interval (−0.04, 0.08). Given the maturity of Scopus, it seems reasonable that there will be few categories for which more than 8% of articles are unciteable for the reason previously outlined, or for which more than approximately 4% of relevant articles with zero-citations are missing from the database. It is therefore reasonable to speculate that if the value of the fitted zero-modification parameter is substantially outside of (−0.04,0.08) then there are doubts about the suitability of the model. The two exceptions to this for the models under consideration here are Health and Social Science and Cultural Studies. For the former, the best fitting model was the

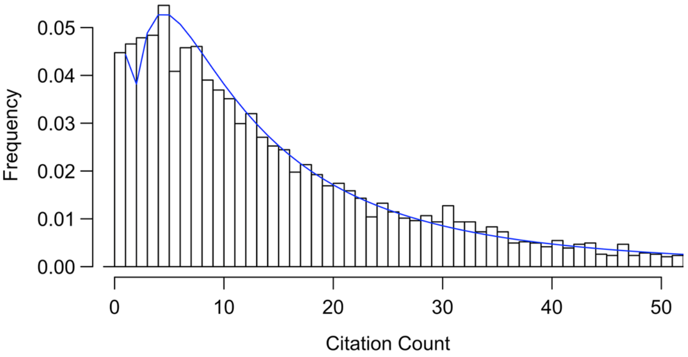

Fig. 5 Example of observed data with non-zero-mode: global and planetary change (best fitted by the zero-modified discrete log-normal in blue). (Color figure online)

zero-modified hooked power law with a zero-modification parameter estimate of 0.16, and for the former the zero-modified Weibull with a zero-modification parameter of − 0.59.

Further insights may be gained by examining the nature of the distributions. As is clear from its probability mass function, the hooked power law distribution is a decaying distribution, that is zeros have the greatest probability of being observed, followed by a 1, followed by a 2 etc., (hence a hooked power law distribution has a zero-mode), the probability of observing a zero being increased if zero-inflation is present. In contrast, a Weibull distribution may have a mode at a positive value. This requires a beta parameter > 1, which is not the case for all such estimates for the data being considered here. A discretised log-normal distribution does in general have a positive mode but may have a zero mode if the values of its two parameters are close. Of course, while zero-inflation may result in a zero-inflated discretised log-normal distribution having a mode at zero, a secondary mode will usually occur at a positive value. Two of the four subject areas for which the discrete log-normal is the best fitting model have an observed mode greater than zero; whilst the observed citation counts for Computer Science Applications have a zero mode, from "Appendix 2" the estimates of the parameters of the non-zero part of the model are 1.56 and 1.29, and hence the data are consistent with a zero-modified discrete log-normal distribution. The remaining subject area that is best fitted by a zero-modified discrete log-normal distribution is Virology, whilst here the observed mode is at zero, there is a secondary mode at 4.

For all other categories the observed citation counts follow a descending pattern and thus are candidates for being best fitted by a zero-modified hooked power law or a zero-modified Weibull, (for Neuropsychology & Physiological Psychology the number of observed 0 s and 1 s are 207 and 213 respectively, but for this subject area there is zero-deflation relative to a hooked power law distribution, and once this deflation has been taken into account the observed distribution decays). For the data considered, in four of the six subject areas for which the zero-modified Weibull is the best fitting distribution, the estimate of the zero-modification parameter is non-significant, the exceptions being Metals and Alloys, which is significant at a level of 0.05, but not at 0.02, and Cultural Studies. Whilst the estimated zero-modification parameter of − 0.594 relative to a zero-modified Weibull distribution is significant, as discussed above such a parameter estimate does not seem feasible. The same is true for the estimates of 0.117 and 0.161 relative to the hooked power law and discrete log-normal, indicating that this subject area should be further investigated.

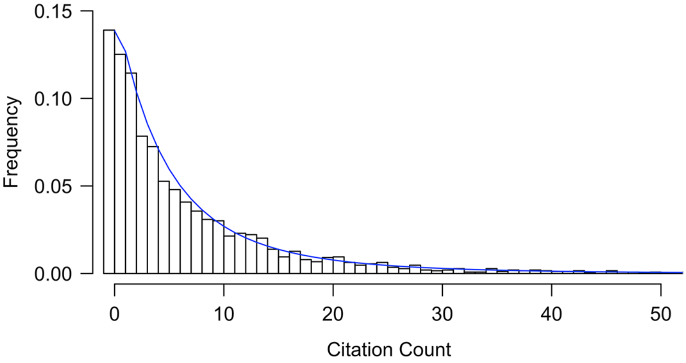

Fig. 6 Example of data with a zero-mode: computational mechanics (best fitted by the zero-modified hooked power law in blue). (Color figure online)

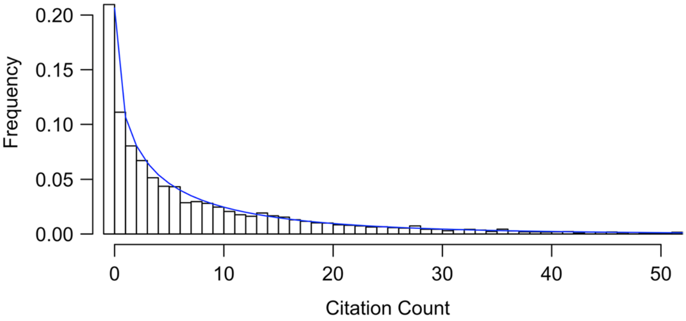

Fig. 7 Example of data with a zero-mode: energy engineering and power technology. Here the number of zeros is much greater than the number of any other count, but there is no evidence of zero inflation relative to a Weibull distribution (best fitted by the zero-modified Weibull in blue). (Color figure online)

Figures 5, 6, and 7 contain examples of the observed distribution of citation counts for subject areas best fitted by zero-modified versions of discrete log-normal, hooked power law, and Weibull distributions.

"Appendix 5" presents (bootstrapped) confidence intervals for the AIC values for the various subject areas under the three distributions considered. If confidence intervals for the AICs of two different models overlap, then there is not significant evidence that one model is better than the other. For any given subject area considered in this paper the confidence intervals of the AICs of all three distributions overlap, and thus it may not be claimed that the "best fitting distribution" fits significantly better than the remaining two. Whilst, say a zero-inflated hooked power law distribution might fit the observed data from a given category best, it is not possible to claim with certainty that this distribution will always be the best fitting model for that category.

Discussion

The principal purpose of this paper is to introduce the concept of zero modified models that admit both zero-deflation and inflation, rather than to discuss the implications of the fitted distributions and estimated parameters for the fields under analysis. Some results are included that are of interest in themselves, however. It is clear from the results that zero-modification is not an absolute concept but occurs relative to a given distribution. For example, the estimated value of the zero-modification parameter for Neuropsychology and Physiological Psychology is 0.044 relative to the discretised log-normal distribution, but – 0.040 relative to a hooked power law distribution, both estimates being significant. Thus, with the former distribution as the base-model there is statistical evidence of zero-inflation and hence "unciteable articles" within the field, but with the latter as the base distribution there is no such evidence of unciteable articles; instead there is evidence that some uncited articles may have been excluded. It is thus important to determine the best fitting base distribution to accurately determine the presence of zero-inflation or zero-deflation (or the absence of either), the presence of zero inflation/deflation relative to one model is insufficient to prove that there are perfect or omitted zeros. It is also important to consider the reality of a model, and not just rely on statistics. If for example an estimate of 0.50 for the zero-inflation parameter occurs, is it feasible that half of the articles in the field under consideration are unciteable?

The zero-modified hooked power law distribution is the best fitting model for 13 subject areas, the zero-modified Weibull best fitting for 6 subject areas, the other 4 being best fitted by the zero-modified discrete log-normal (Table 1). A zero-modified discrete log-normal tends to be the best fitting model when there is a positive mode or secondary mode, the zero-modified Weibull tends to be the best fitting model when there is no evidence of zero-modification relative to this distribution, and the zero-modified hooked power law when the observed citation counts follow a decaying distribution with evidence of zero-modification. These results are not clear-cut, however, and the incorporation of independent variables such as individual, institutional and international collaboration, journal and reference impacts, abstract readability, reference and keyword totals, paper, abstract and title lengths may lead to more precise conclusions. The results comparing distributions are limited to small samples of Scopus categories. Other years and categories may give differing results. The citation count distributions may also be affected by articles published in January having almost a year longer to be cited than articles published in December.

Conclusion

This article introduces zero-modified distributions for citation count data, focussing on zero-modified hooked power law, discrete log-normal and Weibull distributions. The new fitting method allows the estimation of both positive and negative zero-modification parameters, enabling the determination of confidence intervals for and statistical tests of parameter estimates. The results showed that each distribution fits citation count data better than the others for some Scopus categories, and so it seems unlikely that there is a single best distribution for citation count data. The results also show that both zero-inflation and zero-deflation occur for citation count data but changing a base model can alter one type to another. As a consequence of this, it is important to be wary of making definitive statements concerning zero-inflation or zero-deflation. The nature of the distribution of the observed citation counts is also an indicator of the most likely candidate of the mentioned distributions that has the best fit. For cases with the existence of a positive mode or secondary mode, a zero-modified discrete log-normal tends to have the best fit, but for the cases with no evidence of zero-modification relative to the distribution, the zero-modified Weibull is the candidate with the best fit. For the cases with a decaying distribution accompanied by evidence of zero-modification, the zero-modified hooked power law provides the best fit.

Overall, based on the previous research related to the modelling of citation count data, it seems that the non-zero-modified and more zero-modified versions of the mentioned distributions are more compatible with the initial characteristics of citation count data (mass point at zero, highly-right skewness, and heteroskedasticity). The incorporation of independent variables such as individual collaboration, journal internationality, and reference impacts may lead to more precise conclusions.

Appendix 1

See Table 2.

Table 2 The information on the data related to 23 Scopus categories used in this article

Subject	Number of articles	Number of zeros	Percentage of zeros
Food science	4999	799	0.159
Cancer research	5000	712	0.142
Marketing	4762	676	0.141
Filtration and separation	1728	50	0.028
Physical and theoretical chemistry	4999	326	0.065
Computer science application	4999	970	0.194
Management science and operations research	4999	674	0.134
GeoChemistry and petrology	4997	325	0.065
Economics and econometrics	4999	1706	0.341
Energy engineering and power technology	4995	1047	0.209
Computational mechanics	2525	351	0.139
Global and planetary change	3844	172	0.044
Virology	5000	291	0.058
Metals and alloys	4997	1065	0.213
Control and optimization	3059	504	0.164
Critical care and intensive care medicine	4998	917	0.183
Developmental neuroscience	2006	163	0.081
Pharmaceutical science	5000	1122	0.224
Nuclear and high energy physics	4999	644	0.128
Neuropsychology and physiological psych	2827	207	0.073
Health social science	5000	1519	0.303
Cultural studies	5000	2637	0.527
Health information management	993	137	0.137

Appendix 2

See Table 3.

Table 3 Parameter estimates: zero-modified discretised log-normal for 23 Scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	p Value	μ	σ
Food science	0.090	0.0075	0.075	0.105	0.00000	1.82	1.02
Cancer research	0.119	0.0054	0.108	0.13	0.00000	2.43	1.06
Marketing	0.078	0.0071	0.064	0.092	0.00000	1.98	1.10
Filtration and separation	0.021	0.0042	0.013	0.029	0.00000	2.58	0.90
Physical and theoretical chemistry	0.043	0.0040	0.035	0.051	0.00000	2.28	0.94
Computer science application	0.051	0.0114	0.029	0.073	0.00001	1.56	1.29
Management science and operations research	0.089	0.0061	0.077	0.101	0.00000	2.10	1.05
GeoChemistry and petrology	0.043	0.0040	0.035	0.051	0.00000	2.31	0.96
Economics and econometrics	0.190	0.0136	0.163	0.217	0.00000	1.34	1.29
Energy engineering and power technology	0.125	0.0087	0.108	0.142	0.00000	1.82	1.15
Computational mechanics	0.034	0.0126	0.009	0.059	0.00697	1.65	1.06
Global and planetary change	0.027	0.0037	0.020	0.034	0.00000	2.49	1.00
Virology	0.045	0.0035	0.038	0.052	0.00000	2.41	0.91
Metals and alloys	0.116	0.0094	0.098	0.134	0.00000	1.75	1.18
Control and optimization	0.046	0.0129	0.021	0.071	0.00036	1.57	1.08
Critical care and intensive care medicine	0.128	0.0070	0.114	0.142	0.00000	2.16	1.20
Developmental NEUROSCIENCE	0.037	0.0081	0.021	0.053	0.00000	2.18	1.07
Pharmaceutical science	0.128	0.0096	0.109	0.147	0.00000	1.66	1.09
Nuclear and high energy physics	0.076	0.0063	0.064	0.088	0.00000	2.13	1.13
Neuropsychology and physiological psych	0.044	0.0059	0.032	0.056	0.00000	2.21	0.98
Health social science	0.201	0.0108	0.180	0.222	0.00000	1.54	1.07
Cultural studies	0.117	0.0431	0.033	0.201	0.00664	0.10	1.20
Health information management	0.035	0.0193	−0.003	0.073	0.06976	1.83	1.24

Appendix 3

See Table 4.

Table 4 Parameter estimates: zero-modified hooked power law for 23 scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	*p* Value	*B*	α
Food science	0.040	0.0067	0.027	0.053	0.00000	40.21	6.37
Cancer research	0.079	0.0055	0.068	0.090	0.00000	61.35	5.34
Marketing	0.036	0.0064	0.023	0.049	0.00000	27.28	4.18
Filtration and separation	− 0.031	0.0048	− 0.040	− 0.022	0.00000	190.92	12.52
Physical and theoretical chemistry	− 0.014	0.0042	− 0.022	− 0.006	0.00086	121.18	10.90
Computer science application	0.042	0.0084	0.026	0.058	0.00000	11.31	2.96
Management science and operations research	0.042	0.0059	0.030	0.054	0.00000	40.78	5.15
GeoChemistry and petrology	− 0.013	0.0042	− 0.021	− 0.005	0.00197	93.37	8.47
Economics and econometrics	0.191	0.0101	0.171	0.211	0.00000	10.05	3.06
Energy engineering and power technology	0.103	0.0073	0.089	0.117	0.00000	26.56	4.38
Computational mechanics	− 0.018	0.0103	− 0.038	0.002	0.08054	20.28	4.39
Global and planetary change	− 0.022	0.0039	− 0.030	− 0.014	0.00000	79.88	6.37
Virology	− 0.013	0.0039	− 0.021	− 0.005	0.00086	127.17	10.26
Metals and alloys	0.095	0.0076	0.080	0.110	0.00000	19.76	3.77
Control and optimization	− 0.001	0.0102	− 0.021	0.019	0.92190	17.34	4.14
Critical care and intensive care medicine	0.105	0.0064	0.092	0.118	0.00000	31.22	3.86
Developmental neuroscience	− 0.007	0.0075	− 0.022	0.008	0.35065	45.73	5.21
Pharmaceutical science	0.092	0.0080	0.076	0.108	0.00000	23.23	4.66
Nuclear and high energy physics	0.038	0.0058	0.027	0.049	0.00000	31.67	4.15
Neuropsychology and physiological psych	− 0.014	0.0060	− 0.026	− 0.002	0.01963	70.08	7.30
Health social science	0.162	0.0091	0.144	0.180	0.00000	17.89	4.33
Cultural studies	0.161	0.0260	0.110	0.212	0.00000	4.25	3.47
Health information management	0.004	0.0155	− 0.026	0.034	0.79636	13.47	2.94

Appendix 4

See Table 5.

Table 5 Parameter estimates: Zero-modified Weibull for 23 scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	p Value	q	β
Food science	−0.022	0.0121	−0.046	0.002	0.06904	0.82	0.80
Cancer research	0.044	0.0076	0.029	0.059	0.00000	0.90	0.81
Marketing	−0.068	0.0133	−0.094	−0.042	0.00000	0.80	0.70
Filtration and separation	−0.026	0.0064	−0.039	−0.013	0.00005	0.95	1.00
Physical and theoretical chemistry	−0.023	0.0059	−0.035	−0.011	0.00010	0.91	0.94
Computer science application	−0.211	0.0273	−0.265	−0.157	0.00000	0.67	0.56
Management science and operations research	−0.014	0.0098	−0.033	0.005	0.15313	0.85	0.78
GeoChemistry and petrology	−0.029	0.0062	−0.041	−0.017	0.00000	0.91	0.90
Economics and econometrics	−0.036	0.0303	−0.095	0.023	0.23479	0.64	0.57
Energy engineering and power technology	0.008	0.0142	−0.020	0.036	0.57318	0.80	0.72
Computational mechanics	−0.156	0.0256	−0.206	−0.106	0.00000	0.74	0.70
Global and planetary change	−0.044	0.0063	−0.056	−0.032	0.00000	0.91	0.86
Virology	−0.016	0.0052	−0.026	−0.006	0.00209	0.93	0.95
Metals and alloys	−0.040	0.0172	−0.074	−0.006	0.02004	0.76	0.66
Control and optimization	−0.178	0.0275	−0.232	−0.124	0.00000	0.71	0.66
Critical care and intensive care medicine	0.021	0.0113	−0.001	0.043	0.06311	0.83	0.70
Developmental neuroscience	−0.077	0.0146	−0.106	−0.048	0.00000	0.85	0.76
Pharmaceutical science	−0.017	0.0167	−0.050	0.016	0.30870	0.76	0.72
Nuclear and high energy physics	−0.091	0.0126	−0.116	−0.066	0.00000	0.80	0.66
Neuropsychology and physiological psych	−0.040	0.0095	−0.059	−0.021	0.00003	0.89	0.87
Health social science	0.012	0.0211	−0.029	0.053	0.56955	0.70	0.67
Cultural studies	−0.594	0.1712	−0.930	−0.258	0.00052	0.30	0.49
Health information management	−0.229	0.0501	−0.327	−0.131	0.00000	0.70	0.56

Appendix 5

See Table 6.

Table 6 The percentile confidence intervals for AIC values for zero-modified versions of discretised log-normal, hooked power law and Weibull for 23 Scopus categories

Subjects	ZMDLN		ZMHPL		ZMWeibull	
	Lower	Upper	Lower	Upper	Lower	Upper
Food science	31,090.6	31,777.9	31,049.8	31,734	31,053.8	31,761.7
Cancer research	37,077.2	37,860.3	37,027.6	37,810.4	37,117.2	37,909.4
Marketing	32,000.7	32,741.3	31,977.8	32,702.5	32,065.1	32,834.2
Physical and theoretical chemistry	35,866.7	36,448.0	35,858.2	36,450.6	35,904.1	36,513.5
Management science and operations research	34222.9	34,991.7	34,185.5	34,925.6	34,266.9	34,984.1
GeoChemistry and petrology	36,343.0	36,979.4	36,346.6	36,980.8	36,392.2	37076.2
Computational mechanics	15,312.7	15,804.4	15,287.8	15,797.5	15,306.4	15,838.3
Control and optimization	18,104.3	18,677.0	18,066.2	18,665.7	18,125.4	18,719.7
Developmental neuroscience	14,325.6	14,777.2	14,308.3	14,727.7	14,338.1	14,812.5
Nuclear and high energy physics	35,187.9	35,995.2	35,160.9	35,922.2	35,262.2	36,229.3
Neuropsychology and physiological psych	19,998.5	20,493.9	19,953.4	20,464.1	20,016.0	20,514.3
Health social science	26,742.9	27,531.9	26,695.2	27,560.7	26,757.7	27,586.6
Health information management	6573.4	6958.5	6584.9	6969.7	6626.8	6996.6
Economics and econometrics	26,511.9	27,390.0	26,488.6	27,463.9	26,475.0	27,386.7
Energy engineering and power technology	31,445.6	32,256.1	31,443.9	32,194.0	31,344.1	32,166.3
Metals and alloys	31,142.5	31,951.0	31,124.6	31,938.1	31,102.2	31,929.4
Critical care and intensive care medicine	35,038.1	35,855.4	35,007.6	35,837.3	35,012.1	35,834.5
Pharmaceutical science	29,366.2	30,118.5	29,336.4	30,088.5	29,310.2	30,109.5
Cultural studies	16,310.2	17,146.3	16,364.9	17,148.2	16,339.5	17,166.4
Filtration and separation	13,279.3	13,641.4	13,353.5	13,712.8	13,375.9	13,731.0
Computer science application	31,125.3	31,979.7	31,139.7	31,988.3	31,170.8	32,026.9
Global and planetary change	29,679.6	30,260.5	29,702.5	30,280.7	29,771.0	30,370.3
Virology	36,959.5	37,574.0	37,064.3	37,702.5	37,133.6	37,754.5

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from scopus. *Scientometrics, 103*(1), 213–228.

Canty, A., & Ripley, B. D. (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-23.

Dietz, E., & Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics & Data Analysis, 34*(4), 441–459.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Faraway, J. J. (2005). *Extending the linear model with r. generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapmen Hall.

Mendonca, L. (1995). Longitudinalstudie zu kariespraventiven methoden, durchgefuhrt bei 7 bis 10 jahrigen urbanen kindern in belo horizonte (brasilien). Inaugural-Dissertation zur Erlangung der zahnmedizinischen Doktorwurde am Fachbereich Zahn, Mund und Kieferheilkunde der Freien Universitat Berlin.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *PNAS, 99*(8), 5207–5211.

Thelwall, M. (2016). Are there too many uncited articles? Zero-inflated variants of the discretised lognormal and hooked power law distributions. *Journal of Infometrics, 10*(2), 622–633.

Vinciotti, V. (2016). DWreg: Parametric Regression for Discrete Response. R package version 2.0.

Wasserman, L. (2006). *All of nonparametric statistics (springer texts in statistics)*. Berlin: Springer.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters, 127*(C), 51–53.

Wilson, P., & Einbeck, J. (2019). A new and intuitive test for zero modification. *Statistical Modelling, 19*(4), 341–361. https://doi.org/10.1177/1471082X18762277.