Novel statistical models for traditional citations and social media indicators

By: Wan Jing Low (0901253) Supervisors: Prof. Mike Thelwall, Dr. Paul Wilson
Statistical Cybermetrics Research Group, School of Mathematics and Computing, University of Wolverhampton

Introduction

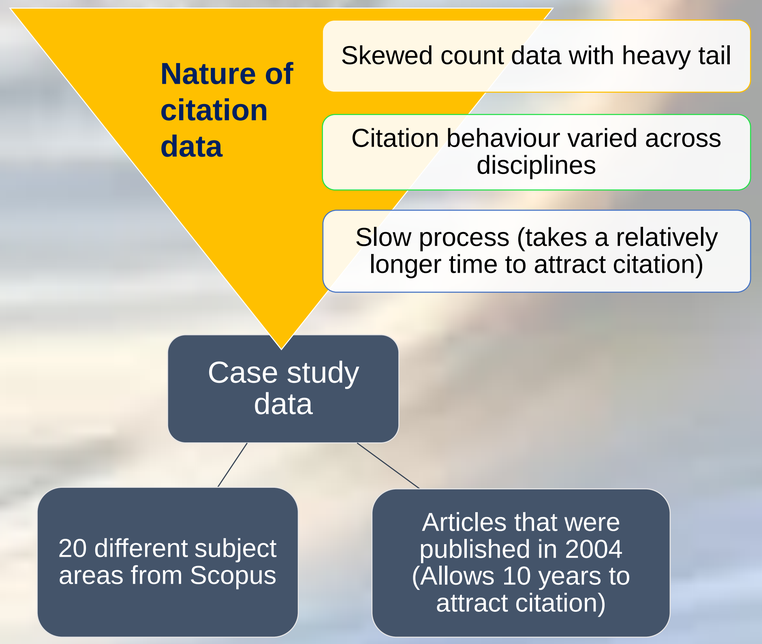
The impact of research publications is of great interest to many scholars. Traditional methods like citation counting have been widely used as indicators for the impact of research and it is therefore important to identify the most appropriate statistical model for citation data to maximise the power of future analyses. Arguably, however, not enough attention has been paid to the selection and validation of appropriate statistical methodologies for quantitative indicators, which is important not only for analyses but also for the understanding of the indicators. Bookstein (2001) pointed out that the measurements in Information Science are often ambiguous, highlighting the challenges faced by any statistical analysis. Hence, it is important to identify the appropriate statistical models to ensure the precision of future predictions. This research aims to assess various statistical models for analysing quantitative indicators for the impact of research outputs, for both traditional and new social media indicators.

Statistical models considered are:

1. Discretised lognormal models
2. Negative binomial (NB) models
3. Stopped sum models
4. Modified stopped sum models

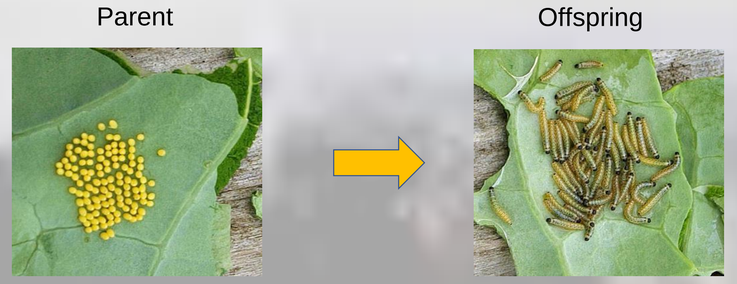
Research Questions

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?



What are stopped sum models?

An example of stopped sum models:



Parent (first wave) population + Offspring (secondary wave) population = Overall stopped sum distribution

Poisson distribution | Negative binomial (NB) distribution | Poisson stopped sum NB distribution

Why stopped sum models are considered?

1. Appropriate for modelling count data
2. Potential to model citation data as two waves, the primary wave and secondary wave
3. The stopped sum models for citation counts could also be appropriate if the two waves occurred simultaneously instead of sequentially

Stopped sum models in citation:

Article gets published in a journal → Article gets cited → Initial citation attracts more citations

Primary (first) wave distribution | Secondary (second) wave distribution

Applied Methodology:

1. Fit case study data with the proposed statistical models using R software
2. Compare models using Akaike Information Criterion (AIC)
   - Estimated by maximum likelihood estimations methods
   - Models with lower AIC are commonly regarded as the better model
3. Determine standard errors for:
   - Negative binomial: obtained directly from R software
   - Discretised lognormal: obtained by bootstrapping
   - Stopped sum models: obtained from the Hessian matrix

Initial results:

1. Stopped sum models especially the modified NB stopped sum NB model produced lower AIC than discretised lognormal
2. However, very large standard errors, hence large confidence intervals were obtained for the parameter estimates of the modified NB stopped sum NB (see Figure 1), indicating the impracticality of the model

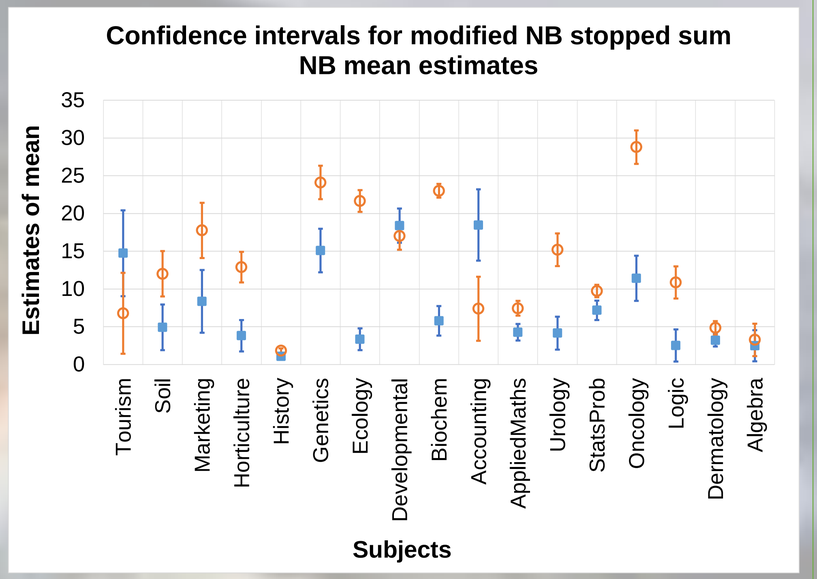


Figure 1: Mean estimates for the modified NB stopped sum NB distribution for both primary (■) and secondary (○) waves with 95% confidence intervals

Initial conclusions:

1. Stopped sum models for citation model are assessed for the first time and they give evidence that there are two processes that influence the citing practices
2. Two 'waves' of citation occur either simultaneously or sequentially
3. Discretised lognormal is more suitable for covariate free data
4. Models with lower AIC may not necessarily be the 'best'

References:

1. 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
2. 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
3. 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/