Application for participation at the Doctoral Forum

Last name: Low

First name: Wan Jing

Institutional affiliation (university/department/link to website):
Statistical Cybermetrics Research Group,
School of Mathematics and Computer Science,
University of Wolverhampton,
Wulfruna Street, Wolverhampton
WV1 1LY
United Kingdom.
Website: http://cybermetrics.wlv.ac.uk/

Full address of the doctoral student including phone and fax numbers and email address:
Flat 10, Birch Court,
Boscobel Crescent,
Wolverhampton.
WV1 1QJ
United Kingdom.
Tel: +447738080137
E-mail: W.J.Low@wlv.ac.uk

Names of the supervisor(s) (links to their websites):
Professor Mike Thelwall (http://www.scit.wlv.ac.uk/~cm1993/mycv.html) and Dr. Paul Wilson

---

(in total max 1500 words)

Description of doctoral research project (including research questions, theoretical background, planned methodology, current status)

Research Questions

a. Which statistical distributions are appropriate for modelling citation data?
b. How can the appropriateness of the proposed models be assessed for citation data? Can these models be extended to altmetric data?
c. What are the issues related to modelling in different disciplines and how can these issues be addressed?

Theoretical background

The use of an appropriate statistical method is important to not only understand the citation process (de Solla Price, 1976), but also to give empirical evidence. Vitanov and Ausloos (2012) argue that statistical approaches could be used to complement qualitative research in scientometrics, allowing knowledge diffusion for better understanding.

Apart from citation, alternative metrics, also known as altmetrics (Priem & Hemminger, 2010) have recently been introduced to complement existing traditional metrics. Altmetrics uses a wide range of data from different platforms including social media platforms and online reference managers. Hence, it will be useful to assess the statistical procedures on non-traditional metrics, such as counts of numbers of readers of articles in the social reference sharing site Mendeley.

A wide range of statistical models have been used previously to model citation data. The typically skewed (de Solla Price, 1976) and heavy-tailed nature of citation data adds difficulty to the task of identifying and fitting appropriate statistical distribution to citation counts. The skewness of citation counts indicate that obtaining the arithmetic mean of citation count or other central tendency statistics will be less useful (Leydesdorff & Milojević, 2012).

General linear models, such as ordinary least squares (Gaussian) regression, have been commonly used to model citation counts (*e.g.* R M Borsuk, Budden, Leimu, Aarssen & Lortie, 2009). However, the general linear model assumes normality for the residuals of its explanatory variables, but this is commonly overlooked in the field, even though a comprehensive guideline on data exploration of citation counts has been given by Bornmann, Mutz, Neuhaus and Daniel (2008).

Until recently, count data regression, such as the Poisson and negative binomial distributions, have been used to model citation data (*e.g.* Didegah & Thelwall, 2013b). In cases where there are excess zeros, the zero inflated negative binomial (ZINB) models have been used (Didegah & Thelwall, 2013a). However, it is difficult to justify the perfect zeros in citations for zero inflation. This is because whilst it is unclear when an academic paper would be cited, fitting a zero-inflated model assumes that there are some papers that due to their nature will never be cited. The hurdle model, which models zeros and positive counts separately may be more appropriate than the zero inflated model, and has also been used to model citation data as it is sensible to assume the presence of a hurdle for a paper to obtain its first citation (Didegah & Thelwall, 2013b).

Planned methodology

a. *Stopped sum distributions for citation data*

Stopped sum models for citation data are assessed for the first time and evidence of two processes influencing citing practises were found. The two processes, which are also known as waves, could occur either sequentially or simultaneously. The two wave interpretation used in stopped sum models could potentially model the 'Matthew effect' in citation (Merton, 1968), if for example, the mean distribution in the second wave is greater than that of the first wave. The assessed stopped sum distributions were fitted to covariate free citation data, it would be advantageous to include covariates into the model to expand its practicality, such as being able to evaluate factors that affect citedness of academic papers, thus revealing the true capability of the model.

b. *Other statistical distributions*

The discretised lognormal distribution has been suggested for citation data from a single subject (Thelwall & Wilson, 2014). However, due to the different citation behaviour across disciplines, it is important to incorporate the wide range of subjects in all analyses.

The negative binomial models have also been fitted previously (*e.g.* Maurseth & Verspagen, 2002). Here the emphasis is on the mean, and the size parameter in the model for estimates are mostly neglected. Modelling this size parameter for estimates in negative binomial will allow us to obtain the variance of an estimated parameter, hence gaining a better understanding of the estimates.

c. *Model selection and validation techniques*

The Akaike Information Criterion (AIC) is a commonly used statistical method for model selection, whereby a model with the lower AIC is commonly being regarded as the better model (Bozdogan, 2000), and a difference of 6 or greater in the AIC would indicate a significant difference between models (Burnham & Anderson, 2003).

According to Vrieze (2012), the AIC may not be consistent in selecting the true model and the AIC may be less efficient if the true model is not within the selection. Therefore further model validation should be carried out when possible, as model selection should not depend solely on AIC. It may be beneficial to determine the standard errors to assess the precision of estimates made by a proposed statistical model. A larger standard error for an estimate would give a larger confidence interval, which would imply that the model is inadequate.

Current status

Initial results have shown that some of the stopped sum models have a relatively lower AIC than the discretised lognormal model, but these models produced very large standard errors, hence large confidence intervals, indicating the imprecision of estimates and impracticality of the model. Therefore, the discretised lognormal model is more suitable for covariate free citation data. Nonetheless, the stopped sum models showed evidence that there are two separate processes that governs citing practices. The initial findings also highlighted the importance of determining standard errors instead of being dependent on the AIC solely.

References

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8,* 93–102. doi:10.3354/esep00084

Borsuk, R. M., Budden, A. E., Leimu, R., Aarssen, L. W., & Lortie, C. J. (2009). The influence of author gender , national language and number of authors on citation rate in ecology. *Ecology, 2,* 25–28. doi:10.2174/1874213000902010025

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.

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. (2013a). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064. doi:10.1002/asi.22806

Didegah, F., & Thelwall, M. (2013b). 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

Leydesdorff, L., & Milojević, S. (2012). Scientometrics. In *The International Encyclopedia of Social and Behavioral Sciences* (pp. 1–20). Retrieved from http://arxiv.org/abs/1208.4566\nhttp://arxiv.org/pdf/1208.4566

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

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science (New York, N.Y.), 159*(3810), 56–63. doi:10.1126/science.159.3810.56

Priem, J., & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, *15*. doi:10.5210/fm.v15i7.2874

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

Vitanov, N. K., & Ausloos, M. R. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion. *Understanding Complex Systems*, 69–125. doi:10.1007/978-3-642-23068-4_3

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*. doi:10.1037/a0027127

Motivation for student participation at the Doctoral Forum and the issues you wish to receive feedback on from the senior researchers.

I am interested in the wide range of topics which would be discussed throughout the doctoral forum. This would definitely be an eye opening experience for me to learn, discover and understand the current challenges and implications in the field. Attending the doctoral forum will also provide a great opportunity for me to meet with experts and fellow doctoral students from all over the world, allowing me to broaden my horizon and further expand my appreciation to the importance of quantitative analysis in various topics.

This forum will also promote opportunities for networking and help to develop my professional skills whilst meeting with people from different background, to learn and understand research from a different perspective. Such event will not only allow the exchange of ideas, but also promote knowledge transfer, thus expanding interdisciplinary knowledge and promoting an innovative research community. An opportunity to take part in the doctoral forum would be extremely beneficial for my professional development whilst enhancing my experience as a PhD student.

The issues I wish to receive feedback from senior researchers are the statistical approaches in various topics in the field, especially in citations and altmetrics. An opportunity to attend the forum would be advantages for me to gain a better understanding on the analyses used whilst gaining an insight from experienced researchers.