Jun 17th, 9:00 AM - 10:20 AM

Predicting citation counts of environmental modelling papers

Barbara J. Robson
CSIRO Land and Water, barbara.robson@csiro.au

Aurélie Mousquès
École des Mines d'Alès, aurelie.mousques-dit-cabanot@mines-ales.org

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Predicting citation counts of environmental modelling papers

Barbara J. Robson[a] **and Aurélie Mousquès** [a,b]
[a] barbara.robson@csiro.au, CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2601, Australia
[b] aurelie.mousques-dit-cabanot@mines-ales.org, École des Mines d'Alès, France

Abstract: We assessed all papers published in two key environmental modelling journals in 2008 to determine the degree to which the citation counts of the papers could be predicted without considering the paper's quality. We applied both random forests and general additive models to predict citation counts using a range of easily quantified or categorised characteristics of the papers as covariates. The more highly cited papers were, on average, longer, had longer reference lists, had more authors, were more likely to have been published in *Environmental Modelling and Software* and less likely to include differential or integral equations than papers with lower citation counts. Other equations had no effect. Although these factors had significant predictive power regardless of which statistical modelling approach was applied, unknown factors (presumably, research quality and relevance) accounted for the majority of variability in citation rates.

***Keywords**:* bibliometrics; h-index; random forest; citations

1 INTRODUCTION

The number of times a paper is cited is a simple metric that is widely used to assess the paper's scientific impact, and is often taken as a proxy for the paper's quality. Citation counts are also the basis for a wide range of other metrics that are increasingly being used (and sometimes misused) to assess the quality of journals and the performance of publishing scientists (e.g. Amez, 2012; Gaster and Gaster, 2012; Kelly and Jennions, 2006; Lazaridis, 2010; Vinkler, 2007).

Citations received by papers have been shown to be influenced by disciplinary domain (Iglesias and Pecharroman, 2007), gender, seniority and stature of the authors (Rossiter, 1993; Slyder et al., 2011; Wu and Wolfram, 2011), prestige of their institution (Wu and Wolfram, 2011), journal of publication (e.g. Slyder et al., 2011), country of residence of authors (Wong and Kokko, 2005), and whether or not the article (Hitchcock, 2013) and the underlying data (Piwowar and Vision, 2013) are available on an open access basis.

Longer papers, especially review articles and others that themselves cite many references, have been found to garner more citations in ecology (Leimu and Koricheva, 2005b), biology (Fawcett and Higginson, 2012), the environmental sciences (Vanclay, 2013) and other fields (Ale Ebrahim et al., 2013; Didegah and Thelwall, 2013).

Several studies (Didegah and Thelwall, 2013; Gazni and Didegah, 2011; Leimu and Koricheva, 2005b) have found that papers with multiple authors are more frequently cited than sole-authored papers, especially when this involves international collaboration (Didegah and Thelwall, 2013; Glanzel, 2001; Leimu and Koricheva, 2005a; Sooryamoorthy, 2009).

Fawcett and Higginson (2012), assessing the citation counts of 649 papers published in leading biology journals in 1998, found that papers with a high density of equations in the main text received fewer citations than other papers: each equation per page in the main text of the paper was associated with a 35% reduction in the number of papers. The authors concluded that equations reduce the accessibility of the paper to a wide readership. This is a finding that may be of concern to modellers, as equations are the tools of our trade. Environmental modelling is a specialised discipline

with a highly numerate population: does this effect hold amongst papers published in journals directed specifically at a modelling readership? This question was one of the motivations of the present paper.

Here, we explore a range of easily quantified or categories attributes of environmental modelling journal papers to evaluate the degree to which they can be used to predict the number of citations the papers had received five years after publication.

2 METHODS

We explored characteristics of all papers published in 2008 in two leading journals that exclusively environmental modelling research: *Environmental Modelling & Software (EMS)* and *Ecological Modelling* (EcoMod). This included a total of 503 papers: 128 published in *Environmental Modelling & Software* plus 375 published in *Ecological Modelling*. The year 2008 was chosen as our reference year, as we considered it recent enough to be relevant to current practise, but sufficiently distant to have allowed differences in citation impacts to emerge.

Each paper was assessed according to a range of quantitative and categorical criteria. The following subset of criteria is considered here:

- **Citation count**: the number of times the paper had been cited at the time of assessment (July-August 2013), as indexed by Web of Science.
- **Page count**: the number of journal pages taken up by the article.
- **Author count**: the number of people sharing authorship credit.
- **Reference count**: the number of articles included in the paper's reference list.
- **Abstract length**: the number of lines in the abstract.
- **Differential equations**: The number of differential, partial differential or integral equations in the main text of the manuscript.
- **Total equations**: the total number of equations in the main text of the manuscript.
- **Journal**: the journal in which the paper was published (EMS or EcoMod).
- **Continent**: The geographic region in which the real-world application is located.
- **Discipline**: the domain of environmental science or modelling to which the paper was most relevant. Each paper was assigned to one of the following categories: aquatic ecology (81 papers), terrestrial ecology (183 papers), theoretical ecology (39), hydrology (27), hydrodynamics (17), water quality (35), meteorology (57), model evaluation techniques (25), uncertainty analysis techniques (5), model visualisation (4), and transdisciplinary (15).
- **Scenarios**: TRUE when the paper describes application of a model to management or change scenarios, FALSE otherwise.
- **Performance metrics**: TRUE when the paper reports quantifications of the model's performance, FALSE otherwise.

These data were used as input variables for a random forest (a more powerful variant of regression trees introduced by Breiman, 2001), using the randomForest package in R (Liaw and Wiener, 2002). This approach was chosen as it very flexible in application to a mix of categorical and numeric predictors with interacting impacts on the response variable, and does not assume a linear response or normally distributed response variable.

An excellent introduction to random forests and other enhanced regression tree methods is given by James et al. (2013), which is available as a free PDF download from the author's website at http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf.

A random forest of 2000 regression trees (more than sufficient to generate stable prediction accuracy) was generated using a randomly selected two-thirds of the sample, drawing from 7 randomly selected variables at each split (following the advice of Liaw and Wiener, 2002), with 3 permutations of out-of-bag observations taken at each iteration to enhance stability. Note that increasing the number of trees in a random forest does not increase the number of parameters in the model, but reduces the parameter (split) estimation error. The one-third of papers not selected for the training subset was used for model testing.

In addition, a generalised additive model (GAM) was developed, using the mgcv package for R (Wood 2011). A GAM is a type of multiple regression model that does not assume a linear response but does assume a normally distributed response variable. An efficient GAM was calculated as

`gam(citations^(1/3)~s(page_count)+s(differential_equations)+other_equations +abstract_length+journal)`. The cube-root transformation of the citation count was necessary to provide an approximately normal distribution for the response variable. *S()* indicates a smoothing term, as described by Wood (2011). Reference count was not included in the GAM as it is highly correlated with page count.

A two-sided Wilcoxon (a.k.a. Mann-Whitney) test (as implemented by Hothorn et al., 2008) was applied to confirm the significance of observed relationships.

3 RESULTS AND DISCUSSION

The length (page count) of a paper and the number of references cited were the two most important factors considered in the random forest (Figure 1). Longer papers and those citing more references are themselves cited more frequently (Figure 2). These are more likely to be review papers or otherwise to be papers that set their results firmly within the context of previous literature and knowledge gaps.

The journal of publication and disciplinary domain also proved important, with papers published in *Environmental Modelling and Software* (EMS) receiving, on average, more citations than those published in *Ecological Modelling* (EcoMod), and papers relating to ecology, hydrodynamics or meteorology receiving significantly fewer citations than cross-cutting papers that discuss model evaluation, decision support systems, or uncertainty analysis (Figure 2). These factors may reflect the wider, less specialised potential audience of EMS and model evaluation papers as well as differences in the editorial policies of the two journals. That EMS has a higher journal impact factor than EcoMod reflects the same underlying relationship.

Continent of application also influenced predicted citation counts (Figure 2): papers relating to applications in Asia, Antarctica or the Arctic Circle received fewer citations than others. Whether or not the model described application to scenarios and whether or not the paper provided metrics of model performance had little effect, though these factors did become marginally significant when a larger random forest, considering a range of additional factors, was generated (Robson and Mousques, submitted).

The final two factors to be discussed (number of authors and number of differential equations) are perhaps the most interesting.

The number of authors was an important factor in the predicted citation count: the more co-authors a paper had, the more often it was cited. Figure 2 shows the shape of this relationship: an increase in citation counts with increasing number of authors up the limit of the data shown (10 authors), though few papers in this sample had more than 6 authors. Part of this effect may be due to an increased opportunity for self-citation and wider exposure of multi-authored papers through the authors' networks, however the contributions of multiple authors to the strength of a paper and in picking up any shortcomings before publication, may also be important.

Papers containing differential, partial differential, or integral equations received fewer citations than those that did not, and the more such equations were included in the paper (up to 5), the fewer citations were predicted (Figure 2).

A larger random forest model, generated using a wider variety of covariates, showed better predictive performance, accounting for 14.5% of the variability in citation counts (Robson and Mousques, submitted). Other covariates considered in this model included the novelty of the model presented in each paper, the application (or not) of the model described to real-world data, the discipline of the application (e.g. meteorology vs. ecology vs. hydrology), the availability of the model software to others, the alphabetical rank of the name of the first author, the number of figures and tables, the length of the title and whether the paper described a new modelling approach or assessment methodology. Those variables discussed in the present paper, however, were the most important variables in the larger model: other variables individually had little influence over the results.

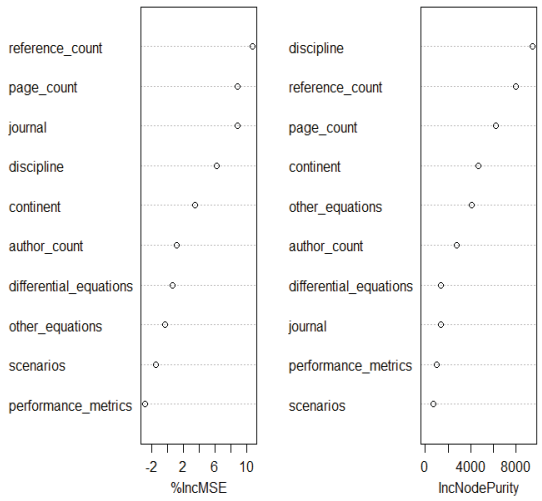

Figure 1. Left: Percent increase in mean square error associated with leaving each predictive variable out of a regression tree. Right: the impact on node purity (a measure of the percentage of misclassified predictions) associated with leaving each variable out of a regression tree. Predictor variables are shown in order of importance.

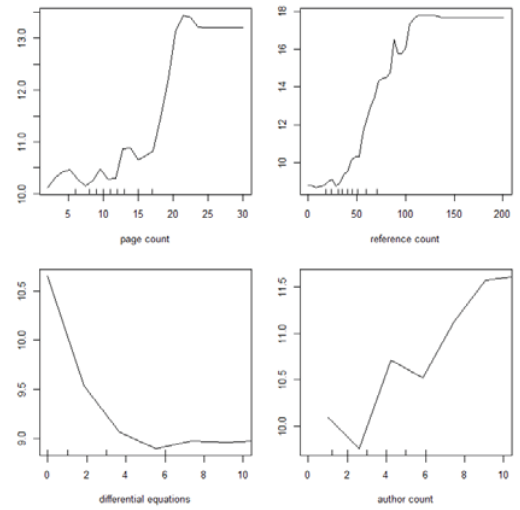

Figure 2. Partial plots from the random forest model, showing the influence of journal discipline, journal of publication, continent of application and whether or not scenarios are simulated on predicted citation counts.

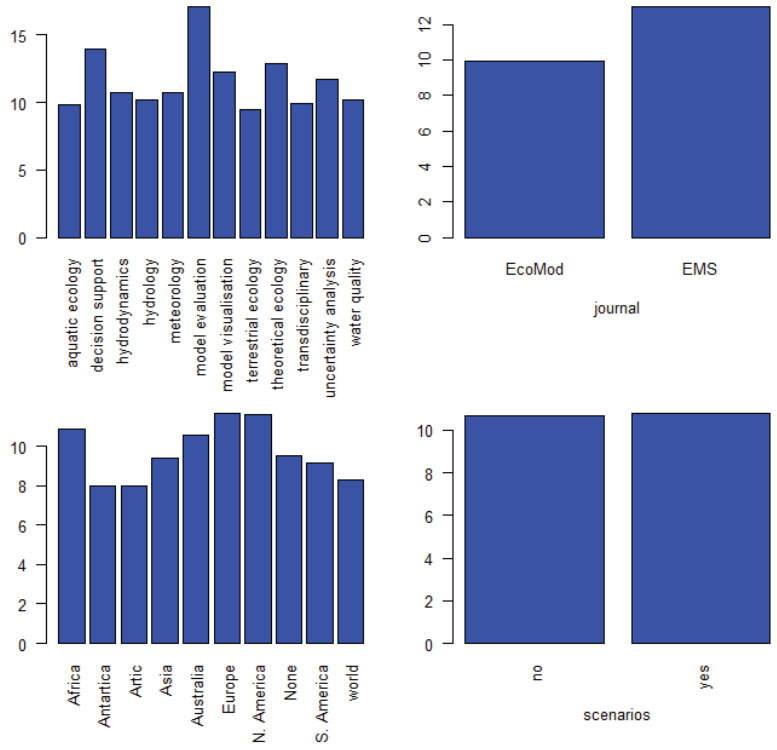

Figure 3. Partial plots showing the marginal effects of page count, reference count, differential equation count and number of authors on predicted citation counts of environmental modelling papers using the random forest model. Interior tick-marks indicate quantiles in the data set.

The GAM approach produced similar results: page count, journal of publication, the number of differential equations, the number of other equations and the length of the abstract were the most significant covariates, combining to produce a model with an adjusted r^2 of 0.25 Unfortunately, however, this model (and all GAM and GLM variants we tried for this dataset) produced heteroscedastic residuals, indicating that it cannot be trusted.

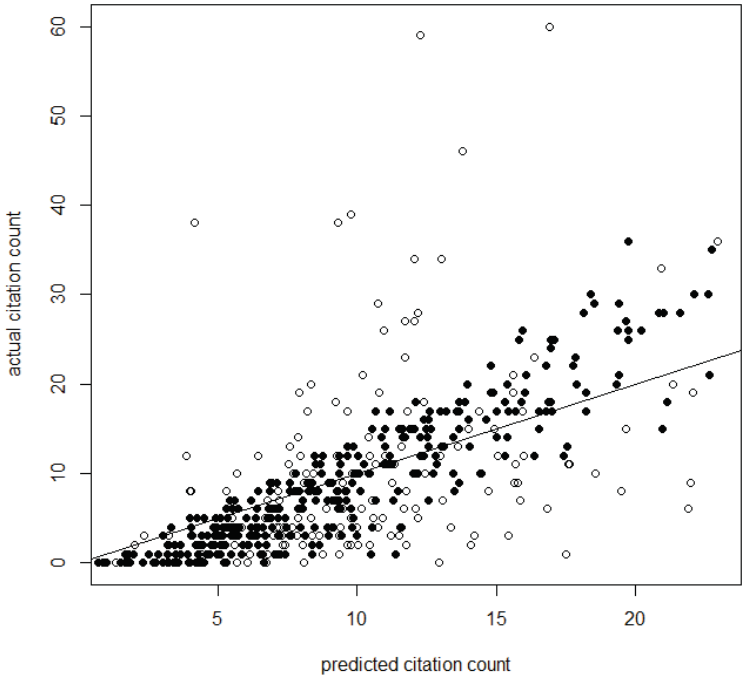

Figure 4. Predicted vs. observed citation count from the random forest model. Solid line indicates a 1:1 relationship. Filled circles indicate samples included in the training data set, open circles include observations held back for testing.

Table 1 Example results of Wilcoxon / Mann-Whitney tests for significance of various statements.

Finding	p-value
Papers with five or more authors had higher citation counts than papers with fewer than five authors	<0.01
Papers reporting models applied to locations in Asia had *lower* citation counts than papers reporting applications in North America, Australia or Europe	<0.05
Papers reporting applications in Antarctica or the Arctic had *lower* citation counts than papers reporting applications Australia or Europe	<0.05
Papers that reported application to scenarios had higher citation counts than papers that did not.	<0.01
Papers that published performance metrics had higher citation counts than papers that did not.	<0.01
Papers of 11 pages or more had higher citation counts than papers of less than 11 pages	<0.001

Papers of between 8 and 10 pages had higher citation counts than papers of less than 8 pages (the shortest 25%)	<0.001
Papers in the top 25% with respect to number of references cited (i.e. those citing 55 or more references) had higher citation counts than papers in the next 25%.	<0.005
Papers in the next quartile (i.e. those citing between 39 and 55 references) had higher citation counts than papers in the following 25% (those citing between 27 and 38 references).	<0.05
Papers in the bottom quartile (i.e. those citing fewer than 27 references) had higher citation counts than those citing between 27 and 38 references.	<0.05
Papers that include no differential or integral equations had higher citation counts than papers that did include differential or integral equations.	<0.001
Papers that included between 1 and 3 differential or integral equations had higher citation counts than papers that included 3 or more differential or integral equations	<0.001

4 CONCLUSIONS AND RECOMMENDATIONS

These results suggest that the number of citations that an environmental modelling journal paper will receive can to some extent be predicted from easily quantified attributes of the paper itself, but that the majority of variation must come from other, less quantifiable factors. This is reassuring, as these factors doubtless include quality of the work, novelty and need for the research, and clarity of presentation.

Paper quality is difficult to assess objectively. Perhaps the most straightforward measure of quality, if available from journal editing offices, would be the score (out of 100) given to each manuscript by reviewers during the refereeing process. Previous work, however, has found only a weak correlation between scores assigned by manuscript assessors and citation counts of papers (Eyre-Walker and Stoletzski, 2013).

Our results show that factors affecting citation counts in the field of environmental modelling are broadly similar to those affecting citation counts in other fields. Authors wishing to maximise citations of their papers may wish to consider publishing longer, more substantive papers on highly generalisable topics, working with multiple collaborators, and limiting the number of differential equations included in the main text of the paper.

A longer version of this paper, considering several additional potential predictive variables and assessing the results in more depth, is in progress.

ACKNOWLEDGMENTS

The second author (Aurélie Mousquès) contributed to this paper as part of an internship undertaken at CSIRO as part of her Bachelor of Engineering degree at the École des Mines d'Alès, France. We thank Dr. Alexander Zwart for the suggestion of using a random forest approach for analysis of our data, and for statistical advice.

REFERENCES

Ale Ebrahim, N., Salehi, H., Embi, M.A., Habibi Tanha, F., Gholizadeh, H., Seyed Mohammad, M., Ordi, A., 2013. Effective strategies for increasing citation frequency. International Education Studies 6(11) 93-99.

Amez, L., 2012. Citation Measures at the Micro Level: Influence of Publication Age, Field, and Uncitedness. Journal of the American Society for Information Science and Technology 63(7) 1459-1465.

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.

Eyre-Walker, A., & Stoletzki, N. (2013). The Assessment of Science: The Relative Merits of Post-Publication Review, the Impact Factor, and the Number of Citations. *PLoS biology*, *11*(10), e1001675.

Fawcett, T.W., Higginson, A.D., 2012. Heavy use of equations impedes communication among biologists. Proceedings of the National Academy of Sciences of the United States of America 109(29) 11735-11739.

Gaster, N., Gaster, M., 2012. A critical assessment of the h-index. Bioessays 34(10) 830-832.

Gazni, A., Didegah, F., 2011. Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. Scientometrics 87(2) 251-265.

Glanzel, W., 2001. National characteristics in international scientific co-authorship relations. Scientometrics 51(1) 69-115.

Hitchcock, S., 2013. The effect of open access and downloads ('hits') on citation impact: a bibliography of studies.

Iglesias, J.E., Pecharroman, C., 2007. Scaling the h-index for different scientific ISI fields. Scientometrics 73(3) 303-320.

James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. Springer DOI 10.1007/978-1-4614-7138-7

Kelly, C.D., Jennions, M.D., 2006. The h index and career assessment by numbers. Trends in Ecology & Evolution 21(4) 167-170.

Lazaridis, T., 2010. Ranking university departments using the mean h-index. Scientometrics 82(2) 211-216.

Leimu, R., Koricheva, J., 2005a. Does scientific collaboration increase the impact of ecological articles? Bioscience 55(5) 438-443.

Leimu, R., Koricheva, J., 2005b. What determines the citation frequency of ecological papers? Trends in Ecology & Evolution 20(1) 28-32.

Piwowar, H.A., Vision, T.J., 2013. Data reuse and the open data citation advantage. PeerJ 1 e175.

Rossiter, M.W., 1993. The Matthew-Matilda Effect in Science. Social Studies of Science 23(2) 325-341.

Slyder, J.B., Stein, B.R., Sams, B.S., Walker, D.M., Beale, B.J., Feldhaus, J.J., Copenheaver, C.A., 2011. Citation pattern and lifespan: a comparison of discipline, institution, and individual. Scientometrics 89(3) 955-966.

Sooryamoorthy, R., 2009. Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. Scientometrics 81(1) 177-193.

Vanclay, J.K., 2013. Factors affecting citation rates in environmental science. Journal of Informetrics 7(2) 265-271.

Vinkler, P., 2007. Eminence of scientists in the light of the h-index and other scientometric indicators. Journal of Information Science 33(4) 481-491.

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society Series B-Statistical Methodology* 73: 3-36. DOI 10.1111/j.1467-9868.2010.00749.x.

Wong, B.B.M., Kokko, H., 2005. Is science as global as we think? Trends in Ecology & Evolution 20(9) 475-476.

Wu, Q., Wolfram, D., 2011. The influence of effects and phenomena on citations: a comparative analysis of four citation perspectives. Scientometrics 89(1) 245-258.