Environmental Modelling & Software 75 (2016) 94–104

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts

Barbara J. Robson[*], Aurélie Mousquès

CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 9 January 2015
Received in revised form
6 October 2015
Accepted 7 October 2015
Available online 11 November 2015

Keywords:
Scientometrics
Informetrics
Bibliometrics
Citation count
Equations

ABSTRACT

We assessed 6122 environmental modelling papers published since 2005 to determine whether the number of citations each paper had received by September 2014 could be predicted with no knowledge of the paper's quality. A random forest was applied, using a range of easily quantified or classified variables as predictors. The 511 papers published in two key journals in 2008 were further analysed to consider additional variables. Papers with no differential equations received more citations. The topic of the paper, number of authors and publication venue were also significant. Ten other factors, some of which have been found significant in other studies, were also considered, but most added little to the predictive power of the models. Collectively, all factors predicted 16–29% of the variation in citation counts, with the remaining variance (the majority) presumably attributable to important subjective factors such as paper quality, clarity and timeliness.

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 (or misused) to assess the quality of journals and the performance of publishing scientists. These include journal impact factors (Garfield, 2006) as well as the now-ubiquitous h-index (Hirsch, 2005), along with a range of alternative indices designed to overcome some of the limitations of the h-index, such as the h(2)-index (Kosmulski, 2006), e-index (Zhang, 2009), the a-, r- and ar-indices (Jin et al., 2007), m-quotient, m-index and h_w-index (Bornmann et al., 2008), g-index (Egghe, 2006), hg-index (Alonso et al., 2010) and $\overline{h}$ (Hirsch, 2010). Many of these indices, their advantages and limitations are discussed in the review by Alonso et al. (2009). Though the inappropriate use of these indices has been widely discussed (e.g. Amez, 2012; Gaster and Gaster, 2012; Kelly and Jennions, 2006; Waltman and Van Eck, 2012), they have been shown to have strong predictive power (Hirsch, 2007) and are now firmly entrenched in academic performance assessment (Alonso et al., 2009; Kelly and Jennions, 2006; Lacasse et al., 2011; Lazaridis, 2010; Oppenheim, 2007; Vinkler, 2007).

Citations received by papers have previously been shown to be influenced by disciplinary domain (Iglesias and Pecharroman, 2007; Slyder et al., 2011), 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 (Didegah and Thelwall, 2013; Judge et al., 2007; Lee et al., 2010; Leimu and Koricheva, 2005b; Slyder et al., 2011), country of residence of authors (Bonitz, 2002; Glanzel, 2001; Leimu and Koricheva, 2005b; 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. Strategies employed by authors to optimise search engine results, such as strategic use of key words and key phrases, making working versions and pre-prints of papers available online, and publicising work through social media, blogs, online tutorials or even email signatures may also have some impact (Ale Ebrahim et al., 2013).

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, 2012a), the environmental sciences (Vanclay, 2013) and other fields (Ale Ebrahim et al., 2013; Didegah and Thelwall, 2013).

[*] Corresponding author.
E-mail address: barbara.robson@csiro.au (B.J. Robson).

http://dx.doi.org/10.1016/j.envsoft.2015.10.007
1364-8152/© 2015 Published by Elsevier Ltd.

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 inter-institutional or (more strongly) international collaboration (Didegah and Thelwall, 2013; Glanzel, 2001; Leimu and Koricheva, 2005a; Sooryamoorthy, 2009). Other studies have not found a relationship between the number of authors and citation counts, but these have tended to be smaller studies, more restricted in scope (Didegah and Thelwall, 2013).

Tregenza (1997) found evidence suggesting that the alphabetical position of an author's first name affected their citation counts.

Leimu and Koricheva (2005b) found that papers in ecology that reported positive findings were more frequently cited than papers reporting negative findings, but found no evidence for any effect of gender, the alphabetical position of the first author's surname, or the journal of publication within this field. Didegah and Thelwall (2013) report that longer abstracts are also associated with enhanced citation counts and that abstract "readability" as assessed by the Flesch Reading Ease Score had no impact on citations.

Fawcett and Higginson (2012a), however, 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 citations. The authors concluded that a high density of equations reduces the accessibility of the paper to a wide readership, and in a subsequent article (Fawcett and Higginson, 2012b), suggested that high equation density is symptomatic of insufficient explanation of theory presented in these papers. 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 study. Here, we endeavour to discover what factors, from amongst those that can be determined without subjective assessment of an article's quality, clarity of presentation, or intellectual significance, can be used to predict the number of citations received by articles published in leading journals specific to environmental modelling.

2. Methods

2.1. Text-mining analysis of >7000 papers

Reference details were downloaded from Web of Knowledge for the 8000 most recently indexed papers published in the following environmental modelling journals: *Ecological Informatics*, *Ecological Modelling*, *Environmental Modelling & Assessment*, *Environmental Modelling & Software*, *Mathematical Modelling of Natural Phenomena* and *Ocean Modelling*. Peer-reviewed and indexed papers published in the proceedings of MODSIM conferences and indexed in Web of Science were also included in the analysis. Duplicate records and those for which no abstracts were provided (primarily comments and prefaces) were excluded, leaving 7602 papers to be considered.

From these reference data, we extracted the following information:

- **Citation count**: the number of times the paper had been cited at the time of assessment (28 September 2014), as indexed by Web of Science. This was converted to the number of citations **per year** since publication. When the exact date of publication was not available, we assumed publication on the 1st of the month, and when the month of publication was not given (23% of cases), we assumed publication on 1 July.

This is the response variable for our model. All those that follow are predictor variables.

- **Year**: The year of publication (516 of the papers included were published in 2005, 777 in 2006, 944 in 2007, 755 in 2008, 794 in 2009, 287 in 2010, 842 in 2011, 708 in 2012, 887 in 2013 and 592 in 2014).
- **Page count**: the number of journal pages taken up by the article.
- **Author count**: the number of authors.
- **Author name**: The position of the first author's name in the dataset, when sorted alphabetically.
- **Journal**: in which journal (or refereed and indexed conference proceedings) the paper was published.
- **Abstract length**: the number of words in the abstract.
- **Title length**: the number of words in the title.
- **Special issue**: whether or not the paper was published as part of a special issue of a journal.

The abstracts were processed using the **tm** (Feinerer et al., 2008) and **topicmodels** (Gruen and Hornik, 2011) packages in R. To simplify analysis, the abstracts were converted to lower case, SMART stopwords (i.e. very common English words) were removed, along with punctuation, numbers, and the following common words: copyright, Elsevier, ltd, rights, reserved, data, can, web, model, models, modeling, modelling, simulation, simulations, level, levels, understanding, developed, effect and effects. The set of abstracts was then processed to:

a) Find lists of words correlated with highly cited (>5 citations per year since publication) or low-cited (<0.5 citations per year since publication, and published prior to 2013) papers, using the "findAssocs()" function from the tm package.
b) Identify "topics" defined by associations of words that tended to appear together, and assign each paper to the topic with which it was mostly strongly associated. For this analysis, we used latent Dirichlect Allocation (LDA; Blei et al., 2003), using the Gibbs fitting method. Each paper was assigned to only one topic, choosing the most likely topic in each case, and setting the topic as "undefined" where the likelihood of the most likely topic was less than twice the probability that the topic was assigned by chance. An optimal number of topics was identified by assessing LDAs that defined between 5 and 20 topics, and choosing the number of topics that resulted in the lowest number of papers with an "undefined" topic.

2.2. Manual analysis of >500 papers

We selected the 511 papers published during 2008 in two journals, *Environmental Modelling & Software* (EMS) and *Ecological Modelling* (EcoMod) for more detailed examination, to allow assessment of a range of additional variables that were not available in the Web of Knowledge reference data. This subset included all 128 papers published in *Environmental Modelling & Software* and all 383 papers published in *Ecological Modelling* in that year, Papers published during 2008 were chosen because we considered 2008 to be recent enough to be relevant to current practice and conditions, but long enough ago that clear differences in citation counts have emerged.

Each paper in this subset was manually assessed according to the following criteria, in addition to those described above:

- **Reference count**: the number of articles included in the paper's reference list.
- **Figure count**: the number of figures per page in the manuscript.

- **Table count**: the number of tables per page in the manuscript.
- **Differential equations**: The number of differential or integral equations per page in the main text of the manuscript.
- **Other equations**: the number of other equations per page in the main text of the manuscript.
- **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 (85 papers), terrestrial ecology (185 papers), theoretical ecology (40), hydrology (27), hydrodynamics (17), water quality (35), meteorology (51), model evaluation techniques (25), uncertainty analysis techniques (5), model visualisation (4), and transdisciplinary (15). For this more detailed analysis, the discipline was determined by manual inspection of the paper rather than relying on topic modelling.
- **Real application**: TRUE (373 cases) if the paper describes an application of a model to a real-world system, FALSE (138 cases) otherwise. This may serve as an indicator of a manuscript's immediate practical relevance.
- **Continent:** The geographic region in which the real-world application is located. Africa (14 cases), Antarctica (4 cases), Asia (54 cases), Australia (22 cases), Europe (161 cases), North America (91 cases), South America (18 cases), world (9 cases) or none (138 cases).
- **Scenarios**: TRUE (145 cases) if the paper describes application of a model to management or change scenarios, FALSE (366 cases) otherwise. This may as an indicator of a manuscript's immediate practical relevance.
- **New model**: TRUE (425 cases) if the paper describes a new model or a substantial development of an existing model, FALSE (86 cases) if the paper describes an application of a previously described model. This may serve as an indicator of manuscript novelty.
- **Novel approach**: TRUE (40 cases) if the paper claims to describe a new approach in modelling or model assessment, FALSE (471 cases) otherwise. This may also serve as an indicator of manuscript novelty.
- **Software availability**: TRUE (232 cases) if the software is available, whether on a commercial or open-source basis, for others to use, FALSE (279 cases) if the software is not available or we were unable to determine the availability of the software.
- **Performance metrics**: TRUE (176 cases) if the paper reports quantifications of the model's performance against validation data, FALSE (335 cases) otherwise. This may serve as one indicator of manuscript quality.
- **DOI date**: The date referenced in the DOI of the article. This is the date at which the DOI was assigned, which may be earlier or later than the publication date, but in some cases may be a better indication of the date from which the paper was first available online.
- **Assessment method**: TRUE (70 cases) if the paper deals primarily with sensitivity analysis, uncertainty analysis, model performance characterisation, model assessment or comparisons between models, FALSE (441 cases) otherwise.

2.3. *Random forest modelling of large (6122 papers) and detailed (511 papers) datasets*

Each of these two datasets was used to provide input variables for a random forest model (a more powerful variant of regression trees; Breiman, 2001) to predict citation counts from the other variables. This approach was chosen as it is 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.

The random forest approach builds a large number of regression or classification trees for the response variable, drawing from a randomly selected subsample of the candidate covariates (predictor variables) at each split (branching of a tree). The importance of each covariate to the final model can be assessed by comparing the average predictive performance of the trees that include a covariate with the average predictive performance of trees that omit that covariate, using bootstrapping to assess predictive performance. If one candidate covariate is correlated with another along an axis relevant to the prediction, the relative importance of each correlated covariate to the performance of the final model will be reduced unless they also have independent predictive power along a second axis. A more complete description and excellent introduction to random forests is given by James et al. (2013).

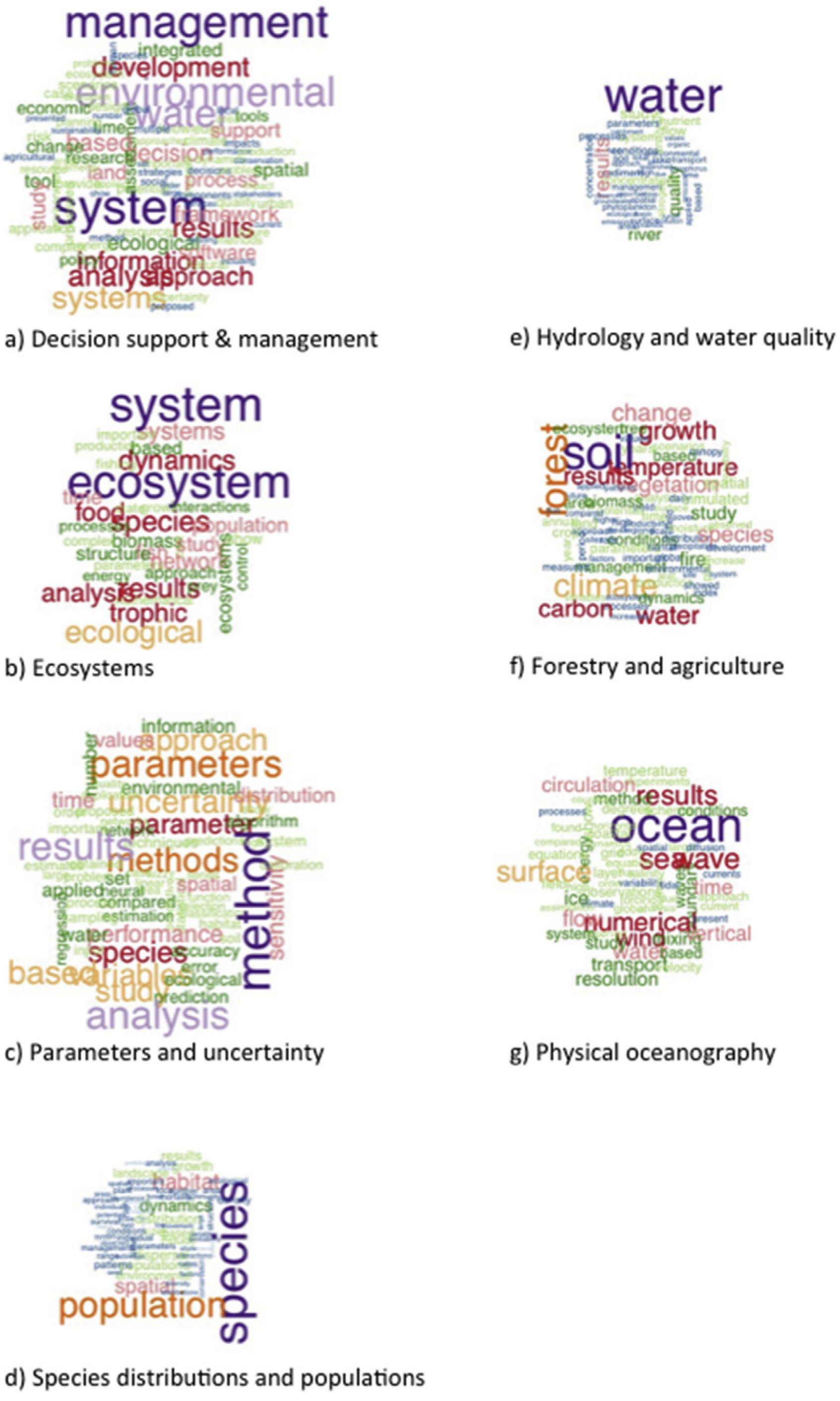
We used the random Forest package in R (Liaw and Wiener, 2002) to complete the analysis, building forests of 500 regression trees (more than sufficient to generate stable prediction accuracy), with 3 permutations of out-of-bag observations taken at each iteration to enhance stability. Out-of-bag observations are a random selection of observations excluded from the dataset for each tree generated. A 10-fold cross-validation was conducted for each random forest to determine the optimal number of covariates (predictor variables) to include.

For the large dataset, we found that the year of publication initially had a disproportionately strong effect on the predicted number of citations per year. Examining the results, it was clear that papers published in the last two years were significantly less cited than older papers, but that the number of publications per year for papers published before 2013 was relatively stable. This may be partly or largely because Web of Science is relatively slow in counting citations derived from recently published papers, and does not include citations in “online first” published in-press publications. To reduce the impact of year of publication on our results, we repeated the analysis, excluding papers published in 2013 or 2014. The random forest model results presented below are for the analysis of the 6122 remaining papers in the large dataset and for the analysis of 508 papers in the smaller dataset.

2.4. *In-depth analysis of citations of top five most-cited papers*

To complement the above analyses, which aimed to predict the citation counts of a broad selection of published papers, we took a closer look at the five most highly cited papers in our dataset with a view to understanding why these papers were being cited. These were:

1. Phillips S.J., Anderson R.P., & Schapire R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, *190*(3), 231–259.
2. Grimm, Volker, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard et al. “A standard protocol for describing individual-based and agent-based models.” *Ecological modelling* 198, no. 1 (2006): 115–126.
3. Austin, Mike. “Species distribution models and ecological theory: a critical assessment and some possible new approaches.” *Ecological modelling* 200, no. 1 (2007): 1–19.
4. Calenge, Clement. “The package “adehabitat” for the R software: a tool for the analysis of space and habitat use by animals.” *Ecological modelling* 197, no. 3 (2006): 516–519.
5. Jakeman, Anthony J., Rebecca A. Letcher, and John P. Norton. “Ten iterative steps in development and evaluation of environmental models.” *Environmental Modelling & Software* 21, no. 5 (2006): 602–614.



a) Decision support & management

e) Hydrology and water quality

b) Ecosystems

f) Forestry and agriculture

c) Parameters and uncertainty

g) Physical oceanography

d) Species distributions and populations

Fig. 1. Commonly occurring words in each of the seven clusters of abstracts identified through topic modelling of the full dataset of 7602 papers. We have assigned the following names to these clusters: a) Decision Support and Management (1423 papers); b) Ecosystems (911 papers); c) Parameters and Uncertainty (1084 papers); d) Species Distributions and Populations (1101 papers); e) Hydrology and Water Quality (887 papers); f) Forestry and Agriculture (905 papers); g) Physical Oceanography (1059 papers). 232 abstracts were unclassified. The size of each word indicates its relative frequency within the cluster of abstracts assigned to that topic. Words occurring in fewer than 10% of abstracts within a topic are not shown.

For each of these primary papers, we took a random sample of 25 citing papers, took note of where the citations occurred (i.e. in the Introduction, Methods, Results, Discussion, Conclusions or Other section of the citing paper) and evaluated the reasons for each citation.

3. Results

From the topic modelling analysis, seven distinct topics emerged. Fig. 1 shows words associated with each of these topics. Increasing the number of topics to ten (not shown) resulted in the emergence of climate change as a distinct topic and the separation of hydrology from water quality modelling and population dynamic

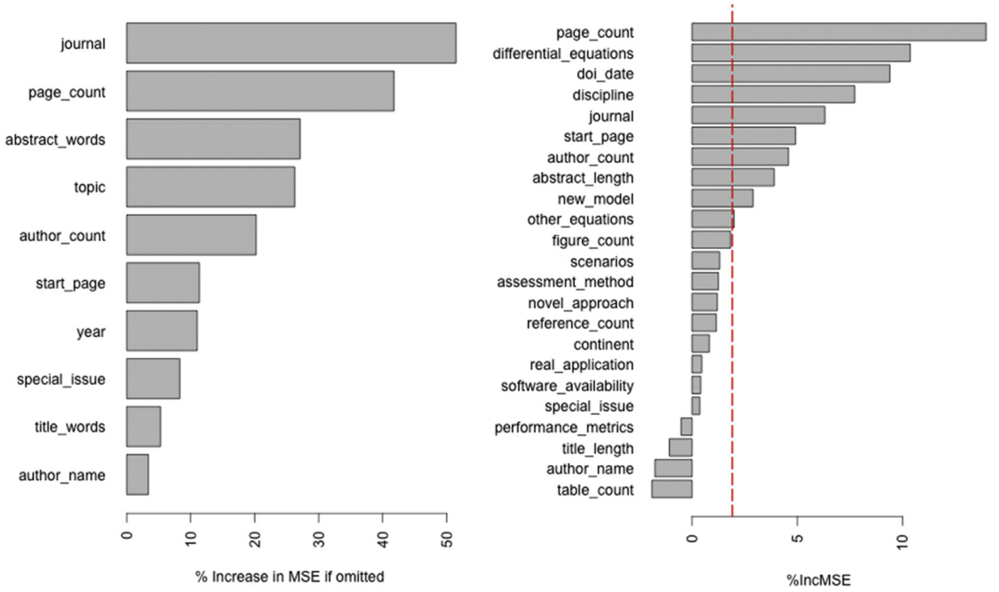


Fig. 2. Variable importance plots for (left) the large dataset (6035 papers published between 2005 and 2012 in nine environmental modelling publications) and (right) the smaller, more detailed dataset (all 511 papers published in *Ecological Modelling* and *Environmental Modelling & Software* in 2008, excluding 9 outliers). The extent of each bar gives the impact on model performance (as indicated by the mean square error of predictions) of omitting the indicated variable from the regression. Variables are sorted in order of importance. In the case of the 2008 dataset, a red dashed line is included to indicate the threshold value below which we can be confident that the impact of the indicated variable is not significant. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

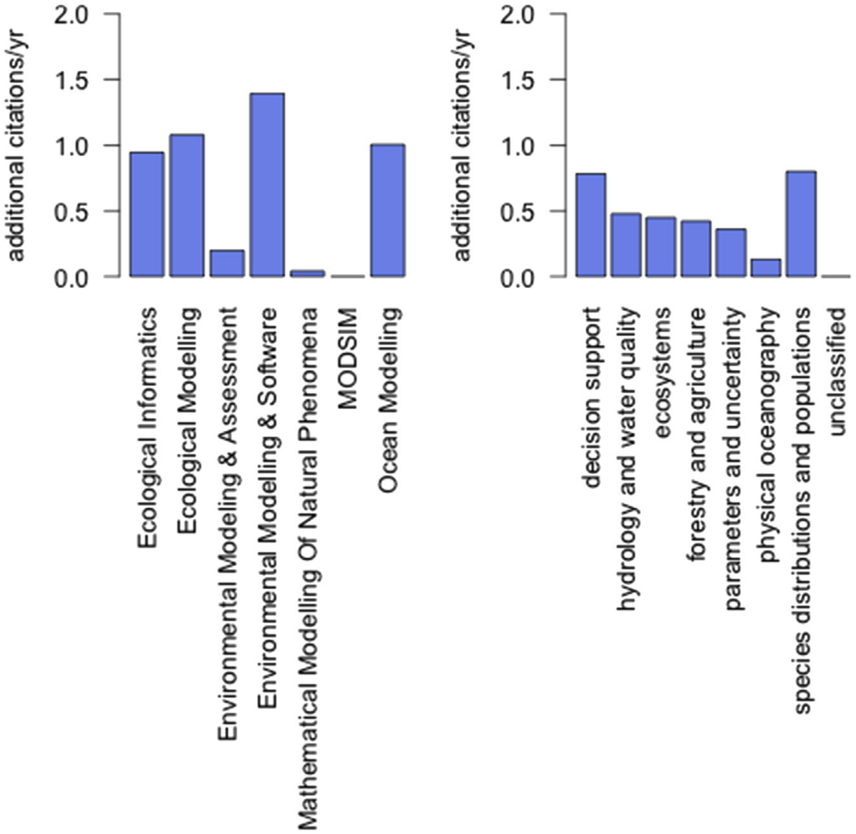


Fig. 3. Normalised partial plots, showing the impact of publication venue (left) and topic (right) on the expected number of citations per year since publication compared with the number expected for the lowest performing option, averaged across all possible values of other variables. Results are from the random forest based on the larger dataset ($n = 6035$), with topics determined automatically. The y-scale is linear and value of 1 indicates that, all else being equal, a paper in that category would be expected to have one more citation than a paper in the worst-performing category. Hence, for example, a paper published in *Environmental Modelling & Software* (EMS) would be expected to have an average of almost 1.5 more citations per year than one published in the MODSIM Proceedings if all other covariates for the two papers were the same.

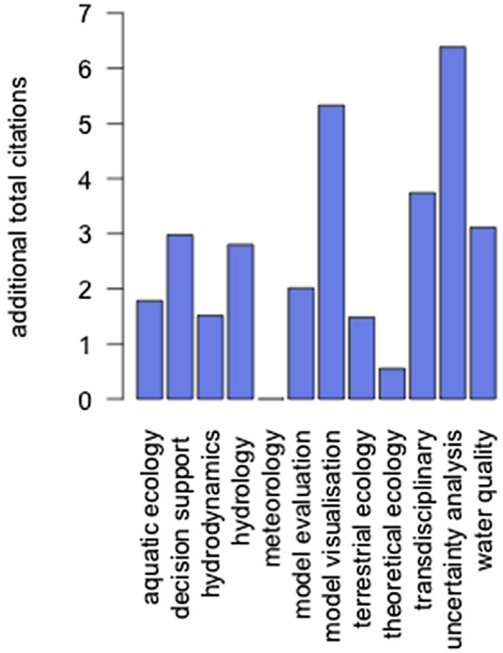


Fig. 4. Normalised partial plot, showing the impact of disciplinary subject area (right) on the expected total number of citations compared with the number expected for the lowest performing option, averaged across all possible values of other variables. Results are from the random forest for the smaller, more detailed 2008 dataset (n = 502), and disciplinary subject area was assessed manually for each paper.

modelling from species distribution and habitat modelling, but reduced the number of papers that could reliably be assigned to any topic.

A small number of very highly cited papers (those with 10 or more citations per year for the large dataset and those with a total of more than 45 citations for the detailed 2008 dataset) appeared to be outliers: omitting these papers from the random forest analyses considerably improved the performance of the random forest models.

The random forest for the larger dataset accounted for approximately 20% of the variability in the citation rates with a mean square error of 2.4 and a correlation between predicted and observed citation rates of 0.52 after very highly cited papers were excluded (leaving 6035 papers in the analysis). This increased to prediction of 29% of the variance (with r = 0.58) if the analysis was restricted to the 3811 papers with low-to moderate total citation counts (total citations less than 10), but fell to only 13% of the variance if outliers were included.

The random forest for the detailed 2008 dataset accounted for approximately 16% of the variability in observed citation counts with a mean square error of 95, after 7 very highly cited papers were excluded. This figure reduced to 8%, if very highly cited outliers were not excluded.

These outlying highly cited papers were more likely to include the following words in their abstracts: **review, overview, applications, global, methods, techniques, approaches, evaluated, practice, guidelines, integrate**. The strengths of these associations were low, with correlations ranging from 0.1 (“review”) to 0.06 (“integrate”).

Papers with low citation counts (fewer than 0.5 citations per year since publication) were more likely to include the words “**mathematical**” or “**figure**” in their abstracts, but again, the association was weak (correlation 0.06).

Results shown in the figures that follow are from random forest models generated while excluding only the very highly cited outliers (n = 6035 for the larger dataset, n = 502 for the detailed 2008 dataset).

Fig. 2 shows the importance of each variable in the random forests. For the larger dataset, the publication venue (journal), paper length (page_count), abstract length, topic and number of authors were important. For the smaller dataset, the results were similar: page count, density of differential equations per page, number of authors, doi_date, discipline and journal of publication dominated the response. Aside from the density of differential and integral equations, the only one of the “extra” variables included in this more detailed analysis that appears to be significant is whether or not the paper purports to present a new model.

After generating Fig. 2, each random forest was recalculated using only the most important covariates (the top 5 in each case) to reduce any potential for over-fitting. This reduced the predictive performance of each model by less that 0.5%.

An advantage of the random forest approach is that it allows us to tease apart cumulative effects of input variables that may be correlated and to examine the shape of the response and graphically identify any break-points. Figs. 3–5 illustrate the impact of each important variable on the predicted citation counts.

Papers published in the established journals – *Environmental Modelling & Software*, *Ecological Modelling*, *Ocean Modelling*, or *Ecological Informatics* – received between 1.2 and 1.7 more citation per year than papers with otherwise similar characteristics published in the MODSIM conference proceedings or in *Mathematical Modelling of Natural Systems*. *Environmental Modelling & Assessment* papers also had lower citation rates (Fig. 3). When considering the smaller, 2008 dataset, a clear difference between papers published in *Ecological Modelling* and *Environmental Modelling & Software* emerged, with the expected number of total citations for papers in papers in *Environmental Modelling & Software* higher by an average of 3.6 compared with the expected number of citations for papers in *Ecological Modelling* with similar characteristics in terms of the other variables considered. This is not unexpected, given the differences in impact factors between the two journals, but demonstrates that the difference is not due to factors such as differences in the average length of papers published in the two journals, the subject domain of the papers, or other characteristics considered in our analysis.

Longer papers with few or no differential or integral equations, longer abstracts and more authors tend to receive more citations. These patterns held true for both the larger dataset (Fig. 5) and the detailed 2008 dataset (Fig. 6). Very short papers (6 pages or fewer) were predicted to receive 0.6 fewer citations per year than papers of more than 12 pages (Fig. 5). There was little additional advantage for papers over 12 pages.

With regard to differential equations, the strongest difference was between papers with no such equations and papers with at least one differential or integral equation – papers with no differential equations are expected to receive on average 4 more citations than papers that do contain differential equations (Fig. 6).

Citation counts increased smoothly with increasing abstract length, whether measured in terms of words (Fig. 5) or lines (Fig. 6), reaching a maximum at about 300 words.

Increasing the number of authors from one to seven increased the predicted citation count by 0.5 per year, further increasing above this so that papers with 12 or more authors are expected to receive 1.0 additional citations per year compared with single-author papers (Fig. 5).

The topic of the paper was also important, with the manually determined disciplinary subject area (Fig. 4) proving to be a much stronger predictor than the automatically assigned topic generated through topic modelling (Fig. 3). Papers relating to uncommon

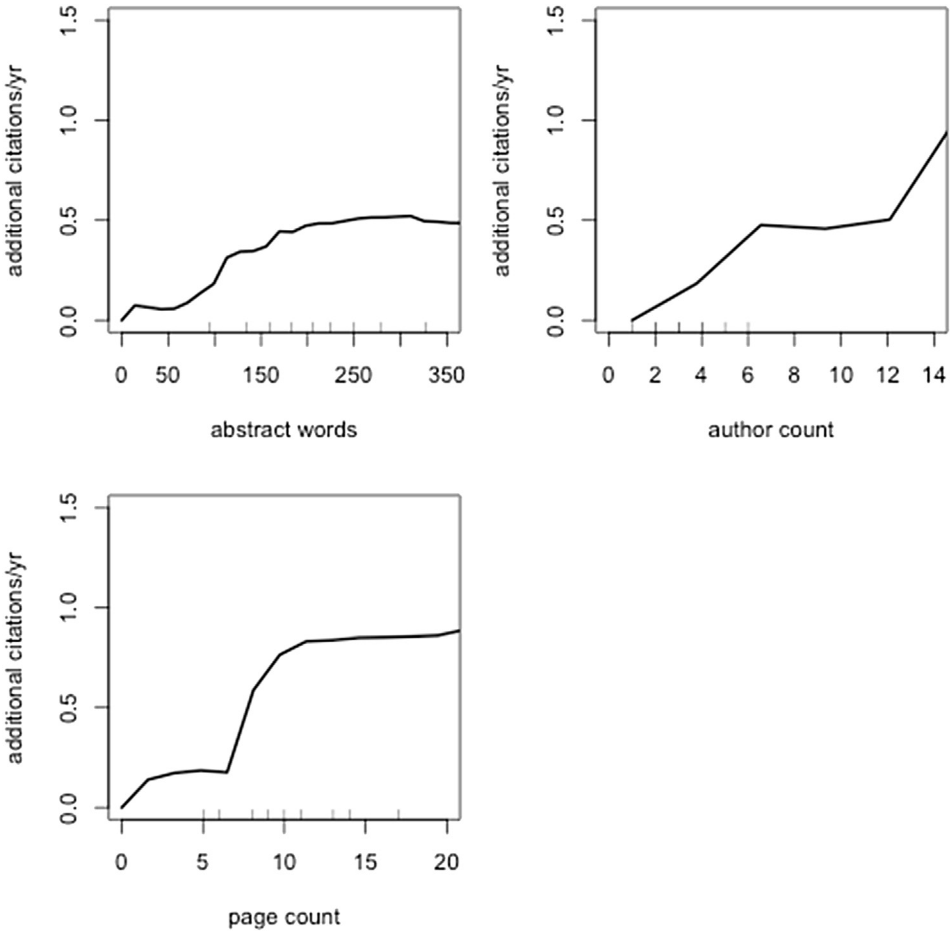


Fig. 5. Normalised partial plots showing the impact of publication year, page count, abstract length and number of authors on the expected total number of citations per year since publication compared with the number expected for the lowest performing option, averaged across all possible values of other variables. Results are for the larger dataset.

topics or subject areas such as meteorology (Fig. 4) and papers that were difficult to assign to a topic (the unclassified papers in Fig. 3) received fewer citations. Papers on topics that cut across disciplinary subject domains, such as uncertainty analysis, model visualisation and transdisciplinary modelling (Fig. 4) and decision support (Fig. 3) are expected to receive the highest citations.

Fig. 7 is presented for comparison with Fig. 5. While Fig. 5 shows the marginal impact of each variable on the expected number of citations per year since publication, Fig. 7 simply shows the mean citation count per year since publication of papers, plotted against the same three variables. This approach exaggerates the impact of author count, page count, and the number of words in the abstract. While a naïve analysis, neglecting the impact of other variables, might suggest that papers with ten or more authors are cited more than three times as often as single-author papers (Fig. 7), the more sophisticated random forest analysis, which adjusts for interactions amongst predictor variables, indicates that a paper with ten or more authors would be expected to have only 0.46 more citations per year than a single-author paper that is similar with respect to other variables such as length, publication venue and topic. The effect of increasing the length of the abstract or the length of the paper are similarly exaggerated in this naïve analysis.

Fig. 8 and Fig. 9 show the results of the analysis of citations of the five most highly-cited papers in our dataset. These very highly cited papers were most commonly cited in the Introduction or Methods sections of citing papers. Every one of the 25 citing papers for Calenge (2006) cited this paper in the Methods section as the source of the software used for analysis. The majority of citations for both Grimm et al. (2006) and Phillips et al. (2006) cite these papers as the source for the method used in the citing papers. Jakeman et al. (2006) and Austin (2007) are cited in a broader range of contexts, most frequently either as references for further discussion of a topic or as support for a statement of fact or opinion that is repeated by the citing paper.

4. Discussion

This analysis suggests that easily quantifiable variables may account for a relatively small (29% at most), but nonetheless significant proportion of the variability in citation counts. More subjective matters such as quality of analyses, clarity of writing and timeliness and level of interest in the subject matter, are likely to be more important. The strength of “journal of publication” as a predictive variable may reflect differences in editorial policies relating to these issues, as well as differences in journal readership and prestige. Journal impact factors are, of course, another way of summarising these effects.

The finding that longer papers received more citations is in accordance with the findings of earlier studies (Leimu and Koricheva, 2005b; Fawcett and Higginson, 2012a; Vanclay, 2013; Ale Ebrahim et al., 2013; Didegah and Thelwall, 2013). If page count is excluded from the analysis, then reference count becomes important, suggesting that these correlated variables are measuring the same effect. Longer papers are more likely to place

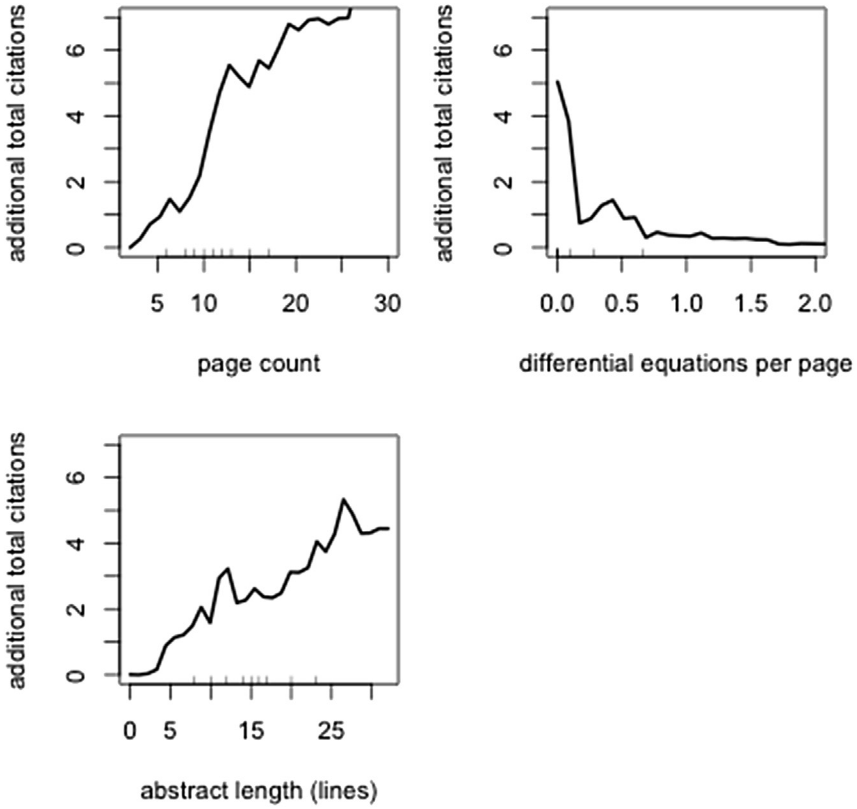


Fig. 6. Normalised partial plots showing the impact of page count, abstract length and density of differential and integral equations on the expected total number of citations per year since publication compared with the number expected for the lowest performing option, averaged across all possible values of other variables. Results are for the smaller, more detailed 2008 dataset.

the work in the context of existing literature and are more likely to be review papers that are useful references for students and later authors. The relatively high citation rates of review papers is a well-known phenomenon, and is also reflected in the list of words associated with very highly cited papers in this analysis. Longer papers may also contain, on average, more substantive content.

Papers that include differential or integral equations have received fewer citations than other papers, echoing the findings of Fawcett and Higginson (2012a) for more general biological literature. This result is not a small effect, and holds regardless of whether we use the total number of equations or the number of equations per page as the relevant covariate in our random forest model. Given the highly numerate readership of the journals considered in our analysis, how is this to be interpreted? One possibility is that equations do reduce the accessibility of the work, reducing the readership to more highly motivated readers. Another possibility is that papers with differential equations appeal to more specialised (and therefore smaller) audiences, as they are less likely to be review articles, position papers, commentaries, or applications of models that are already well-known and widely used. That most of the effect is found in the difference between papers with no differential or integral equations and papers with one or more such equations and there is very little additional penalty for including more than one equation (Fig. 6) might tend to support the latter explanation.

Interestingly, neither the density nor the total count of other equations (statistical equations and simple algebraic equations) had a significant effect.

Papers relating to topics that cut across disciplinary boundaries of model application, such as model evaluation frameworks, uncertainty assessments, and decision support, have received slightly more citations than others. This may reflect the wider potential readership of these papers.

Papers relating to topics in ecology or meteorology received, on average, fewer citations than those relating to hydrology, agriculture or forestry. It is possible that the journals included in this review are less widely read in the ecological and meteorological sciences than in some of the other fields considered, or this may reflect differences in citation practices between fields.

In this analysis, unlike those of other authors (Bonitz, 2002; Glanzel, 2001; Leimu and Koricheva, 2005b; Wong and Kokko, 2005), geographic location was not an important factor in predicting citation counts. Note, however, that we considered the location of the model application rather than the addresses of the authors, and we considered this factor only in our smaller sample, so this result does not directly contradict the findings of previous authors.

We hypothesised that the alphabetic position of the first author's name would influence position in search engine results, and hence citation counts, but found no evidence to support this hypothesis. If there was such an association in 1997 (Tregenza, 1997), perhaps improvements in search engines have removed this effect. The number of co-authors, however, does make a small difference, with increasing citations associated with increasing author counts. Some part of this association may be due to increased opportunities for self-citation (e.g. Robson, 2005), but working with co-authors also offers the chance to widen the context of the work, drawing on a wider pool of expertise, while providing more perspectives on the relevance and key findings of the work, and more opportunity for collaborative revision and critical review of the manuscript

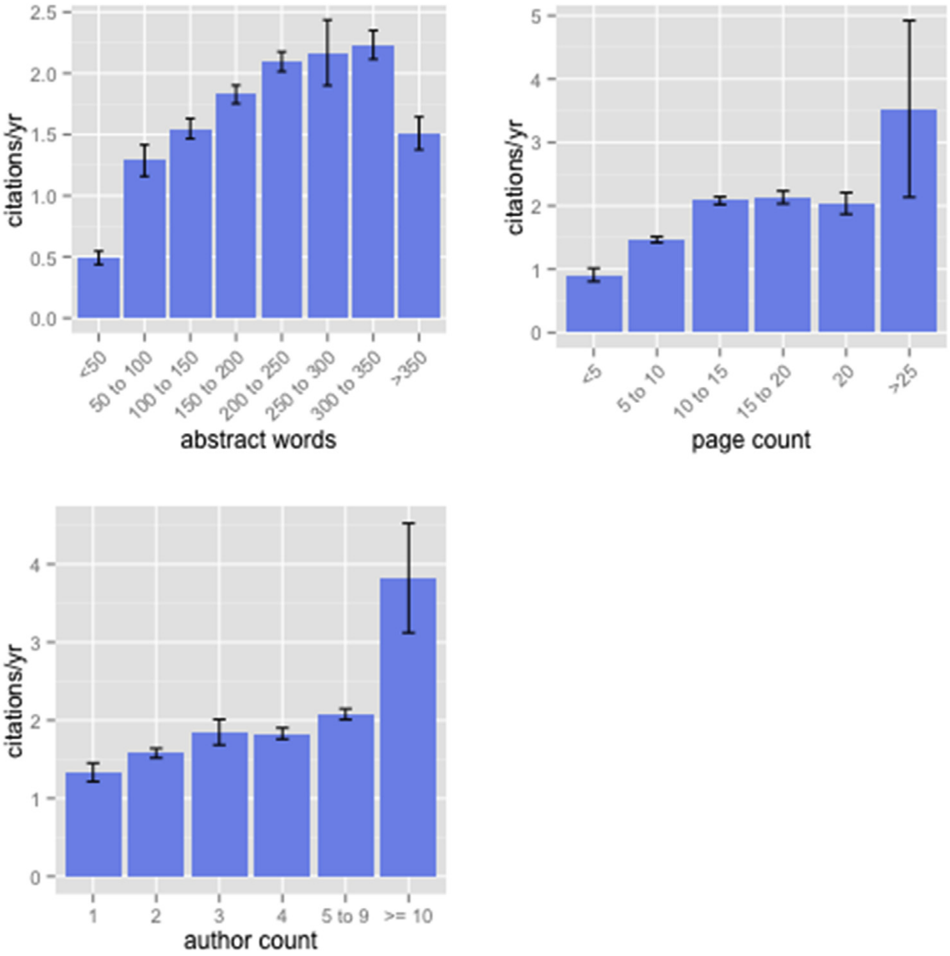


Fig. 7. Raw citation count per year since publication of papers in the larger dataset plotted against four variables, ignoring interactions between covariables. Error bars show the standard error.

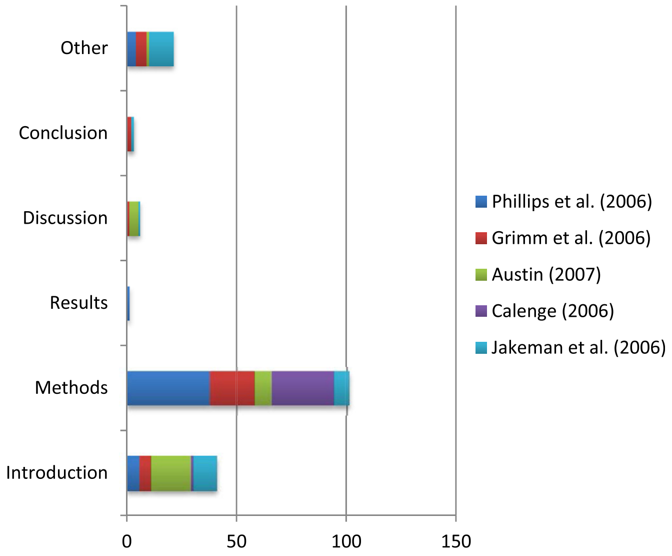


Fig. 8. Summary of where in citing papers each of the five most highly-cited papers in our dataset were cited. Results are from a random sample of 25 citing papers for each of the five highly cited papers. Values indicate the total number of citations within this sample, colours indicate the cited paper. Some citing papers cite the source paper more than once. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

before it is submitted.

The most highly cited papers in our dataset were cited for a range of reasons, but fell into three categories: papers that introduced a new method that was subsequently widely used (Phillips et al., 2006; Grimm et al., 2006), papers that provided a tool or software that could be used by others to simplify their analyses (Calenge, 2006; Phillips et al., 2006), and papers that provided a critical review or synthesis that helped others to structure their work or put it into context (Austin, 2007; Jakeman et al., 2006).

Many papers (nearly 8% of our >500 paper detailed dataset) attempt or purport to introduce a new method, but in general, this is not predictive of subsequent impact as indicated by citation counts (Fig. 2). It is not easy to develop a new approach that will transform established practise. Those that do successfully introduce a new method that is widely adopted, however, may become very highly cited outliers.

Many papers are based on software that is available to others in some form, but again, this is not generally predictive of citation counts (Fig. 2). Calenge (2006) made available an easy-to-use software tool for an in-demand method (also introduced by Calenge in an earlier and less frequently cited paper), and distributed it as an R toolkit. R is both widely used and encourages and facilitates citation of sources by its users. The lesson here: make it easy for others to use your methods and make it easy for them to cite you, and you are more likely to have success.

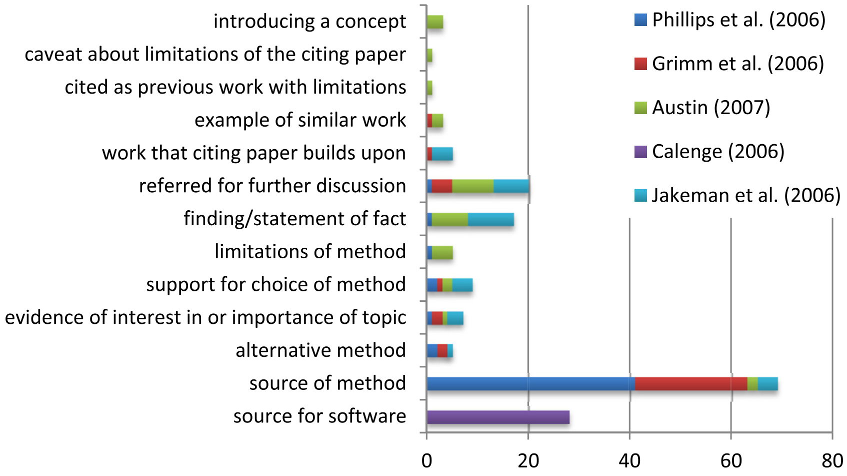


Fig. 9. Summary of reasons for citation of each of the five most highly cited papers in our dataset. Results are from a random sample of 25 citing papers for each of the five highly cited papers. Values indicate the total number of citations within this sample, colours indicate the cited paper. Some citing papers cite the source paper more than once. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

5. Conclusion

Although the impact of all of these quantifiable factors together is relatively small and the relationships identified are correlative, not necessarily causative, some general recommendations can be made if authors wish to maximise the number of citations each of their papers receives. Whether maximising citations per paper is preferable to maximising an author's total number of citations across many papers is open to debate and may influence the decision of whether to aim for longer, more substantial papers (which do receive more citations per paper) or smaller, "least publishable unit" papers.

To maximise citations per paper, authors should, where practicable:

- Place their work firmly within the context of other relevant work and knowledge gaps in the literature, even if this results in a longer manuscript: longer manuscripts receive, on average, more citations that short manuscripts (Maier, 2013 offers advice on writing an effective literature review).
- Ensure that any equations used are explained as clearly and simply as possible, or else consider whether they can be replaced with plain English descriptions and relegated to an appendix with more detailed explanation. Papers containing differential or integral equations tend to receive fewer citations than those without, and this might be partly because they are less accessible. It may also, however, simply be an indicator of a paper targeted at a narrower and more specialised audience, in which case, this advice may not help.
- Make it as easy as possible for others to apply their methods, for instance through distributing software toolkits or providing a step-by-step process, following the examples of the very highly cited papers that we examined (e.g. Calenge, 2006).
- Clearly explain how the work is relevant beyond a narrow disciplinary domain, especially if the paper belongs to one of the less well-cited disciplines within the environmental modelling literature such as ecology or meteorology; and
- Work collaboratively, drawing on the knowledge and expertise of co-authors to improve the quality and general relevance of their papers. Papers with more authors receive, on average, more citations, though the effect is small in our dataset when adjusted for other factors.

Author contributions

Ms Mousques collated the 2008 data, assessed normative attributes such as the novelty of the model presented by each paper, and conducted preliminary statistical analyses, while Dr Robson applied the random forests and text mining techniques, assessed reasons for citations, and wrote the bulk of this paper.

Acknowledgements

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.

This manuscript builds substantially on a more constrained previous analysis of the 2008 dataset that has been published as a conference paper (Robson and Mousques, 2014).

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. Int. Educ. Stud. 6 (11), 93–99.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2009. h-Index: a review focused in its variants, computation and standardization for different scientific fields. J. Inf. 3 (4), 273–289.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2010. hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices. Scientometrics 82 (2), 391–400.

Amez, L., 2012. Citation measures at the micro level: influence of publication age, field, and uncitedness. J. Am. Soc. Inf. Sci. Technol. 63 (7), 1459–1465.

Austin, Mike, 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200 (1), 1–19.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bonitz, M., 2002. Ranking of nations and heightened competition in Matthew core journals: two faces of the Matthew effect for countries. Libr. Trends 50 (3), 440–460.

Bornmann, L., Mutz, R., Daniel, H.D., 2008. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. J. Am. Soc. Inf. Sci. Technol. 59 (5), 830–837.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Calenge, Clement. The package "adehabitat" for the R software: a tool for the analysis of space and habitat use by animals. Ecol. Model. 197 (3), 2006, 516–519.

Didegah, F., Thelwall, M., 2013. Which factors help authors produce the highest impact research? collaboration, journal and document properties. J. Inf. 7 (4), 861–873.

Egghe, L., 2006. An improvement of the h-index: the g-index. ISSI Newsl. 2 (1), 8–9.

Fawcett, T.W., Higginson, A.D., 2012a. Heavy use of equations impedes communication among biologists. Proc. Natl. Acad. Sci. U. S. A. 109 (29), 11735–11739.

Fawcett, T.W., Higginson, A.D., 2012b. Reply to Chitnis and Smith, Fernandes, Gibbons, and Kane: communicating theory effectively requires more explanation, not fewer equations. Proc. Natl. Acad. Sci. 109 (45), E3058–E3059.

Feinerer, Ingo, Hornik, Kurt, Meyer, David, 2008. Text mining infrastructure in R. J. Stat. Softw. 25 (5), 1–54. http://www.jstatsoft.org/v25/i05/.

Garfield, E., 2006. The history and meaning of the journal impact factor. JAMA-Journal Am. Med. Assoc. 295 (1), 90–93.

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.

Grimm, Volker, Berger, Uta, Bastiansen, Finn, Eliassen, Sigrunn, Ginot, Vincent, Giske, Jarl, Goss-Custard, John, et al., 2006. A standard protocol for describing individual-based and agent-based models. Ecol. Model. 198 (1), 115–126.

Gruen, Bettina, Hornik, Kurt, 2011. topicmodels: an R package for fitting topic models. J. Stat. Softw. 40 (13), 1–30. http://www.jstatsoft.org/v40/i13/.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A. 102 (46), 16569–16572.

Hirsch, J.E., 2007. Does the h index have predictive power? Proc. Natl. Acad. Sci. U. S. A. 104 (49), 19193–19198.

Hirsch, J.E., 2010. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics 85 (3), 741–754.

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.

Jakeman, Anthony J., Letcher, Rebecca A., Norton, John P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21 (5), 602–614.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York, pp. 303–332.

Jin, B.H., Liang, L.M., Rousseau, R., Egghe, L., 2007. The R- and AR-indices: complementing the h-index. Chin. Sci. Bull. 52 (6), 855–863.

Judge, T.A., Cable, D.M., Colbert, A.E., Rynes, S.L., 2007. What causes a management article to be cited – article, author, or journal? Acad. Manag. J. 50 (3), 491–506.

Kelly, C.D., Jennions, M.D., 2006. The h index and career assessment by numbers. Trends Ecol. Evol. 21 (4), 167–170.

Kosmulski, M., 2006. A new Hirsch-type index saves time and works equally well as the original h-index. ISSI Newsl. 2 (3), 4–6.

Lacasse, J.R., Hodge, D.R., Bean, K.F., 2011. Evaluating the productivity of social work scholars using the h-index. Res. Soc. Work Pract. 21 (5), 599–607.

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

Lee, S.Y., Lee, S., Jun, S.H., 2010. Author and article characteristics, journal quality and citation in economic research. Appl. Econ. Lett. 17 (17), 1697–1701.

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 Ecol. Evol. 20 (1), 28–32.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2 (3), 18–22.

Maier, H.R., 2013. What constitutes a good literature review and why does its quality matter? Environ. Model. Softw. 43, 3–4.

Oppenheim, C., 2007. Using the h-index to rank influential British researchers in information science and librarianship. J. Am. Soc. Inf. Sci. Technol. 58 (2), 297–301.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190 (3), 231–259.

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

Robson, B.J., 2005. Representing the effects of diurnal variations in light on primary production on a seasonal time-scale. Ecol. Model. 186 (3), 358–365.

Robson, B.J., Mousques, A., 2014. Predicting citation counts of environmental modelling papers. In: Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.), Proceedings of the 7th International Congress on Environmental Modelling and Software. International Environmental Modelling and Software Society (iEMSs), San Diego. June 2014.

Rossiter, M.W., 1993. The Matthew-Matilda effect in science. Soc. Stud. Sci. 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.

Tregenza, Tom, 1997. Darwin a better name than Wallace? Nature 385 (6616), 480–480.

Vanclay, J.K., 2013. Factors affecting citation rates in environmental science. J. Inf. 7 (2), 265–271.

Vinkler, P., 2007. Eminence of scientists in the light of the h-index and other scientometric indicators. J. Inf. Sci. 33 (4), 481–491.

Waltman, L., Van Eck, N.J., 2012. The inconsistency of the h-index. J. Am. Soc. Inf. Sci. Technol. 63 (2), 406–415.

Wong, B.B.M., Kokko, H., 2005. Is science as global as we think? Trends Ecol. Evol. 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.

Zhang, C.T., 2009. The e-index, complementing the h-index for excess citations. PLoS One 4 (5).