palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 19 Feb 2015 | Accepted 28 Apr 2015 | Published 2 Jun 2015

DOI: 10.1057/palcomms.2015.11

OPEN

Analysis of bibliometric indicators to determine citation bias

Ivan Simko[1]

ABSTRACT Citations of research papers and citation-related indicators are frequently used factors in determining research priorities, allocating funding, and deciding appointments, promotions and tenures. The main problem with using citations for a variety of evaluations is a substantial difference in the average number of citations received by papers in different research fields and subfields. A large number of species are subjects of biological research, but the distributions of their citations have not been studied in detail. The key objective of the present work was to determine whether the choice of experimental subjects influence bias in citations. A case study was performed on papers from 108 plant species and five research fields. Funnel plot analyses and computer simulations identified species-related citation bias within all research fields. Relationships between bibliometric indicators imply that species with a fast growing number of publications in recent years (for example, new model organisms) generally have a higher average number of citations per paper than is the overall mean for the research field. In contrast, the average number of citations received by the five most prominent papers of the species was strongly correlated with the total number of published papers from laboratories working with the species. The current study indicates that despite the high frequency of cross-species citations, citations of species show a pattern similar to separate subfields. Although these analyses were performed on plant research papers, the findings have relevance for other areas of research where experimental subjects tend to form separate subfields. To reliably compare citations across subfields, a new type of bibliometric indicator is needed. In the meantime committees evaluating quality of research should take into consideration that citations of papers within a research field might be biased due to the experimental subjects used in the studies.

[1] U.S. Department of Agriculture, Agricultural Research Service, U.S. Agricultural Research Station, Salinas, CA, USA

Introduction

Assessing the importance and quality of research is a complex and challenging task that is traditionally performed by a panel of experts familiar with the research field. Although the expert-based evaluations are generally well accepted, the process is slow and depends on random elements such as the selection of experts (Cole *et al.*, 1981). Several methods have been proposed to alleviate these problems using bibliometric indicators that are based on the number of published papers and their citations (Hirsch, 2005; Jin *et al.*, 2007; Bornmann *et al.*, 2008). Despite certain shortcomings, objections and caveats (Seglen, 1997; Campbell, 2008; Reinstein *et al.*, 2011), citations and citation-related indicators are regularly applied to assess the scientific impact of journals (Garfield, 1972), strengths of research groups (Schubert *et al.*, 1989) and performance of individual scientists (Hirsch, 2005). Citations are also utilized together with other criteria to determine research priorities (Neff and Corley, 2009), allocate funding, and decide appointments, promotions and tenures (Reed, 1995; Ball, 2007). The main problem with using citations for evaluations is a large difference in the average number of citations received by papers in different research fields (Seglen, 1997; Iglesias and Pecharromán, 2007; Radicchi *et al.*, 2008). Therefore measurements of performance that are derived from citation count cannot be directly compared across research fields. While analyses of bibliometric indicators were performed previously for multiple research fields (Iglesias and Pecharromán, 2007; Radicchi *et al.*, 2008; Althouse *et al.*, 2009) and subfields (Narin *et al.*, 1976; Vinkler, 1988), they were never studied at the species level. A large number of species are used as experimental subjects in many areas of biological research, including plant and animal science, mycology, bacteriology, and others. In contrast to relatively well-separated fields and subfields that were investigated previously (Narin *et al.*, 1976; Vinkler, 1988), the use of a species is not limited to a specific research field. The same species can be used as an experimental subject in several research fields but often at a different frequency within each. The identification of the citation pattern at the species level is needed to develop more objective approaches for comparing scientists' performances. This article describes a case study performed on plant species within specific research fields; however, the findings have relevance for the analysis of citations in other areas of academic research.

Anecdotal evidence suggests that research performed on certain plant species is more likely to be cited. Plant researchers who do not work with Arabidopsis occasionally, half-jokingly mutter that one can grow a single Arabidopsis plant on a windowsill and publish a highly cited paper from a 6-month study, but it is hard to get the same number of citations by publishing similar research performed on crop species tested in multiple environments for years. This statement is a deliberate oversimplification that mixes types of research, study durations, needs for replication and other aspects. It raises, however, an interesting question about the possible effect of plant species used in research studies on the citation of published papers.

The citations of research papers show a highly skewed distribution with few papers having many citations (Seglen, 1992; Redner, 2005; Lundberg, 2007). Differences in the number of citations can be observed among papers, but also in the average number of citations received by research groups (for example, departments, institutions and countries), research journals or research fields (Glänzel, 1996; Tijssen *et al.*, 2002; Lundberg, 2007; Adler *et al.*, 2009). To compare citations of papers at the plant species level, analyses were performed on data obtained from two databases for plant scientific publications, Science Citation Index Expanded and Scopus. Data have been acquired for research papers published in the 10-year period from 1995 throughout 2004, and cited from 1995 throughout July 2013. Plant research fields covered in this study are Genetics & Heredity (G&H), Physiology & Biology (P&B), Pharmacology, Toxicology & Pharmaceutics (PTP), Plant Pathology (Path) and papers published in a Crop Science journal (CSw).

The objectives of the present work are to investigate the current status and trends in citations of plant research publications and to compare citation patterns at the plant species level. Relationships between bibliometric indicators were examined to determine whether the number of published papers and changes in species popularity are associated with the citation pattern of species. Knowledge about citations and bibliometric indicators can be taken into consideration when making decisions about research needs or comparing research performances.

Methodology

Database search. Searches were carried out in the Science Citation Index Expanded database of Web of Science (WoS) (Thomson Reuters, New York, NY, USA) and in the Scopus (SCO) (Elsevier, Amsterdam, the Netherlands) database in July 2013. Both databases were used in this study because each of them provides automatic grouping of journals for a different research field: WoS for "G&H" and SCO for "PTP". Data have been obtained for research papers published in the 1995–2004 period. This 10-year period was chosen because papers published during this period already have a sufficient number of citations for reliable statistical analysis, and because the rate of accruing new citations is relatively small compared to already accumulated citations, indicating the steady state of citation distribution (Stringer *et al.*, 2008). Only regular research papers published in journals were analysed. Searches in databases were performed using scientific names of species of the kingdom *Plantae* according to the International Plant Names Index (http://www.ipni.org), the USDA Plants Database (http://plants.usda.gov), the USDA Germplasm Resources Information Network (http://www.ars-grin.gov) and the Encyclopedia of Life (http://eol.org) accessed in July 2013 (Supplementary Table S1). All but one analysed species belong to subkingdom of vascular plants (*Tracheobionta*). Searches were performed with accepted scientific names and frequently used synonyms indicated in the four databases. A single exception to this rule was triticale, a hybrid of two species, for which combinations of the scientific and the common names were used due to the fact that the scientific name is seldom provided in publications. Searches were performed for 130 species covering a broad range of plants; however, only species with 10 or more papers published in the 1995–2004 period were statistically analysed. Raw data for all citations are publicly available from the Harvard Dataverse repository (Simko, 2015). For the 2005 to July 2013 period the number of papers but not their citations were recorded; these data were used to compare frequency of species-related papers in two time periods. Searches of scientific names of species were executed in publications titles, abstracts and keywords. Five research fields analysed in the present work cover different sets of journals:

G&H: The WoS database was searched with the species' scientific name used in the field "Topic". The resulting list was then limited to articles in the "Genetics Heredity" research field.

PTP: The SCO database was searched with the species' scientific name used in the field "Article Title, Abstract, Keywords". The resulting list was limited to PTP "Subject Area".

P&B: The SCO database was searched with the species' scientific name used in the field "Article Title, Abstract, Keywords", and "ISSN" codes for 33 plant physiology and biology journals (Supplementary Table S2). These journals were selected from the "Plant Science" and the "Forestry" subject

categories in the SCImago database (http://www.scimagojr.com). The selection of journals was based on three criteria: a SCImago Journal Rank indicator above 0.3 in 2011, the journal not being included in any other analysed research field and the author's personal classification of journals into research fields.

Path: The SCO database was searched with the species' scientific name used in the field "Article Title, Abstract, Keywords", and "ISSN" codes for 19 plant pathology journals (Supplementary Table S2). These journals were selected from the "Agronomy" and the "Forestry" subject categories listed in the Journal Citation Report database (Thomson Reuters, New York, NY, USA). The other three selection criteria were the same as for the P&B category.

CSw: These searches were performed to compare results from a single journal that covers a relatively narrow research field to those obtained for wider research fields with numerous journals. The WoS database was searched with species scientific name used in the field "Topic", and Crop Science in the "Publication Name" field. The results of CSw searches were visually inspected and the papers describing registration of a new germplasm or cultivar were deleted from the list.

Bibliometric indicators. Citations of published papers always show skewed distribution (Seglen, 1992; Redner, 2005; Lundberg, 2007); therefore a logarithmic transformation was performed to achieve normality of data distribution. All logarithmic transformations in the present work were a natural logarithm. To avoid the problem of zero values on the logarithmic scale a value of 1 was added to the number of citations that each paper received. This adjustment changes the number of citations to "less than" category (Seglen, 1992); for example, articles with zero citations belong to the category with less than one citation. Bibliometric indicators used in this article follow acronyms, denotations and calculations of Karolinska Institute (Rehn *et al.*, 2007). In addition, new indicators were introduced that allow analyses of citation data performed in the present study (Table 1).

Statistical analyses. Statistical analyses were performed on field-normalized data (Lundberg, 2007; Rehn *et al.*, 2007). Citation *z*-score averages ($\bar{c}_{fz[\ln]}$) within each research field were compared using Funnel plot analysis (Spiegelhalter, 2005). This simple graphical approach allows detection of plant species with values significantly different from the overall mean. The threshold to declare the difference significant was set to correspond to the 95% control limits of the overall mean value adjusted for the number of published papers and analysed species. The same approach was used to compare differences among citation z_{T5}-score averages for *T*5 ($\bar{c}_{fz[\ln]-T5}$). Statistical analyses of a number of publications for a single species (P_{s95-04}) and a relative growth rate (RGR) in the share of publications were performed using one-sample and two-sample *z*-tests for proportions, respectively. Relationships between bibliometric indicators were analysed using the Pearson correlation coefficient and tests of association. Simulated citation z_{T5}-score averages for *T*5 ($\hat{c}_{fz[\ln]-T5}$) and their 95% confidence intervals were determined from 20,000 datasets that have been generated by randomly shuffling citation data within each research field.

The Pearson correlation coefficient and tests of association were performed in JMP 11.1.1 (SAS Institute, Cary, NC, USA), Funnel plot analysis and *z*-test for proportions were carried out in Microsoft Excel v.14.1.4 (Microsoft, Redmond, WA, USA). *P*-values for multiple comparisons were adjusted by the false discovery rate approach (Benjamini and Hochberg, 1995) to keep the experiment-wise threshold at $\alpha = 0.05$. The code for reshuffling citation data and calculating simulated $\hat{c}_{fz[\ln]-T5}$ scores was written in the Python programming language.

Results

Average number of citations. Database searches of the five research fields yielded 75 (G&H), 100 (P&B), 78 (PTP), 70 (Path) and 27 (CSw) species with 10 or more papers published during the 1995–2004 period. The largest number of publications for a species in research fields was 1583 in G&H for *Arabidopsis thaliana*, 2857 in P&B for *Arabidopsis thaliana*, 796 in PTP for

Table 1 | Bibliometric indicators, their acronyms, denotations and calculations

Acronym	Description
P_{95-04}	Number of publications for a research field published in 1995–2004 period
P_{05-13}	Number of publications for a research field published in 2005–2013 period
P_{s95-04}	Number of publications for a single species published in a research field during 1995–2004 period
P_{s05-13}	Number of publications for a single species published in a research field during 2005–2013 period
p_{s95-04}	Relative share of publications for a single species published in a research field during 1995–2004 period; $p_{s95-04} = P_{s95-04}/P_{95-04}$
p_{s05-13}	Relative share of publications for a single species published in a research field during 2005–2013 period; $p_{s05-13} = P_{s05-13}/P_{05-13}$
RGR	Relative growth rate in the share of publications for a species; $RGR = \ln(p_{s05-13}/p_{s95-04})$; this indicator is relating the publication growth rate of a species to the growth rate in contemporaneous publications from the same research field; positive values of RGR identify species with a higher grow rate between two periods than is the average grow rate for the research field
*T*5	A group of five most highly cited (prominent) publications of a species in a research field
$c_{fz[\ln]}$	Item oriented field normalized logarithm-based citation *z*-score for a single publication *i* (or citation *z*-score); $c_{fz[\ln]} = \left(\ln(c_i + 1) - \left[\mu_{f[\ln]}\right]_i\right) / \left[\sigma_{f[\ln]}\right]_i$ where c_i is number of citations to publication *i*; $[\mu_{f[\ln]}]_i$ is the average value of the logarithm of the number of citations plus one to publications from the same research field as publication *i*; and $[\sigma_{f[\ln]}]_i$ is the standard deviation of the $[\mu_{f[\ln]}]_i$ distribution; citation *z*-score expresses how far a value is from the population mean in terms of the number of standard deviations (Lundberg, 2007, Rehn *et al.*, 2007)
$\bar{c}_{fz[\ln]}$	Item oriented field normalized logarithm-based citation *z*-score average (or citation *z*-score average); $\bar{c}_{fz[\ln]} = \frac{1}{P}\sum_{i=1}^{P}\left(\left(\ln(c_i + 1) - \left[\mu_{f[\ln]}\right]_i\right) / \left[\sigma_{f[\ln]}\right]_i\right)$ where *P* is the number of publications for which the average is calculated (in this work $P = P_{s95-04}$); note that the acronym uses *c* with overbar; (Lundberg, 2007, Rehn *et al.*, 2007)
$\bar{c}_{fz[\ln]-T5}$	Item oriented field normalized logarithm-based citation z_{T5}-score average for *T*5 (or citation z_{T5}-score average); the score is calculated as citation *z*-score average, but from five most highly cited publications only. Note that the acronym uses *c* with overbar
$\hat{c}_{fz[\ln]-T5}$	Simulated citation z_{T5}-score average for *T*5 (or simulated citation z_{T5}-score average); averages and confidence intervals of $\hat{c}_{fz[\ln]-T5}$ are calculated from simulated datasets that have been generated by randomly shuffling citation data 20,000 times. Note that the acronym uses *c* with hat
$r\bar{c}_{fz[\ln]-T5}$	Residual value of citation z_{T5}-score average for *T*5 (or residual of citation z_{T5}-score average) is calculated by subtracting the simulated citation z_{T5}-score average for *T*5 from the observed citation z_{T5}-score average for *T*5; $r\bar{c}_{fz[\ln]-T5} = \bar{c}_{fz[\ln]-T5} - \hat{c}_{fz[\ln]-T5}$. Positive values indicate that observed scores are higher than simulated scores while negative values indicate that observed scores are lower than simulated scores. Note that the acronym uses c with overbar

Zea mays, 633 in Path for *Solanum lycopersicum* and 364 in CSw for *Zea mays* (Supplementary Table S3). Citations of published papers in all research fields showed lognormal distribution. Because citation *z*-scores ($c_{fz[\ln]}$) were calculated from logarithmically transformed citation data, their distributions were close to normal (Figure 1).

The citation *z*-scores averages ($\bar{c}_{fz[\ln]}$) ranged from −1.29 for *Hevea brasiliensis* in G&H to 0.96 for *Medicago truncatula* in Path (Supplementary Table S3). Funnel plot analysis that takes into the consideration the number of publications per species and the number of species per research field was used to identify plant species with $\bar{c}_{fz[\ln]}$ significantly different from the overall mean (Figure 2). Highly significant differences were detected among species in all research fields. Nineteen species had average citation scores significantly different from the overall mean in two or more research fields. Five of these species (*Arabidopsis thaliana, Medicago truncatula, Oryza sativa, Solanum lycopersicum* and *Vitis vinifera*) had significant $\bar{c}_{fz[\ln]}$ always higher than the overall mean indicating that these species are cited more frequently than expected. Ten species (*Avena sativa, Beta vulgaris, Digitalis lanata, Helianthus annuus, Hevea brasiliensis, Linum usitatissimum, Medicago sativa, Pisum sativum, Secale cereale* and *Triticosecale rimpaui*) had significant $\bar{c}_{fz[\ln]}$ always lower than the overall mean indicating that these species are cited less frequently than expected. The remaining four species (*Hordeum vulgare, Nicotiana tabacum, Phaseolus vulgaris* and *Triticum aestivum*) showed mixed results signaling that citations of these species depend on the research field. To study relationship between citation *z*-scores averages ($\bar{c}_{fz[\ln]}$) and the number of published papers for species (P_{s95-04}) in more detail, correlations between the two bibliometric indicators were calculated within research fields. Correlations ranged from $r = 0.080$ (G&H) to $r = 0.344$ CSw, with weak, but significant correlations detected in the fields of P&B ($r = 0.285$, $P = 0.004$) and Path ($r = 0.312$, $P = 0.008$). These data show a positive relationship between P_{s95-04} and $\bar{c}_{fz[\ln]}$, but the lack of significant correlation across all research fields indicates that the number of publications alone cannot simply explain the differences in citation *z*-scores averages.

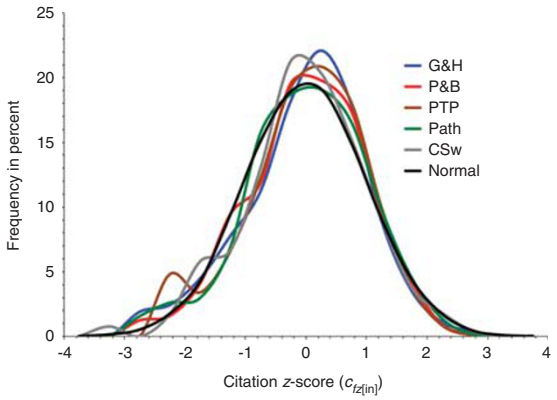


Figure 1 | Distribution of citation *z*-scores ($c_{fz[\ln]}$) received by the papers published from 1995 throughout 2004.
***Notes*: Abbreviations for the research fields are: G&H for Genetics & Heredity, P&B for Physiology & Biology, PTP for Pharmacology, Toxicology & Pharmaceutics, Path for Plant Pathology, and CSw for papers published in Crop Science journal. The black line shows normal distribution with a mean of 0 and a standard deviation of 1.**

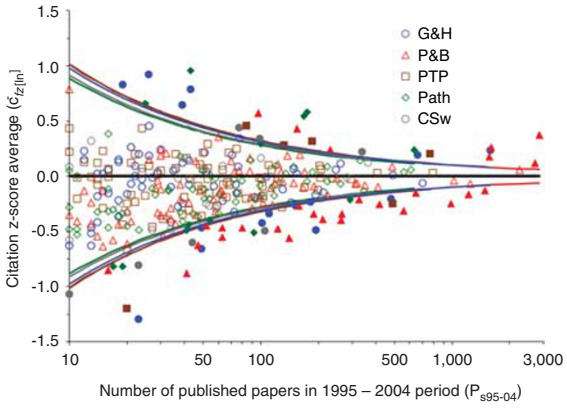


Figure 2 | Funnel plot analysis of the citation z-score averages ($\bar{c}_{fz[\ln]}$) for 108 plant species and five research fields.
***Notes*: The black horizontal line indicates the overall mean of 0, while coloured lines indicate the 95% control limits of the overall mean. The $\bar{c}_{fz[\ln]}$ values outside the control limits (fully coloured symbols) are significantly different from the overall mean. The X-axis is in a logarithmic scale. Abbreviations for the research fields are the same as in Figure 1. Values of $\bar{c}_{fz[\ln]}$ for species and research fields are in Supplementary Table S3.**

Citations of most prominent papers. Differences in citation z_{T5}-score averages ($\bar{c}_{fz[\ln]-T5}$) among plant species were highly significant in all research fields and ranged from −0.23 for *Nicotiana tabacum* in CSw to 3.62 for *Arabidopsis thaliana* in P&B (Supplementary Table S3). The $\bar{c}_{fz[\ln]-T5}$ scores showed a strong and significant ($P < 0.0001$) correlation with the logarithm of P_{s95-04} (Figure 3) in all research fields (ranging from $r = 0.808$ in G&H to $r = 0.895$ in P&B) demonstrating that citation z_{T5}-score averages grow with increasing number of published papers. A positive correlation between the two bibliometric indicators has been anticipated because *T*5 papers represent different proportions of publications for species with different number of published papers. For example, five papers represent 50% of all publications in species with 10 published papers; however, the percentage drops to 0.5% for species with 1000 published papers. Therefore, if citations of papers were randomly distributed, the citation *z*-score average ($\bar{c}_{fz[\ln]}$) would be expected to be similar

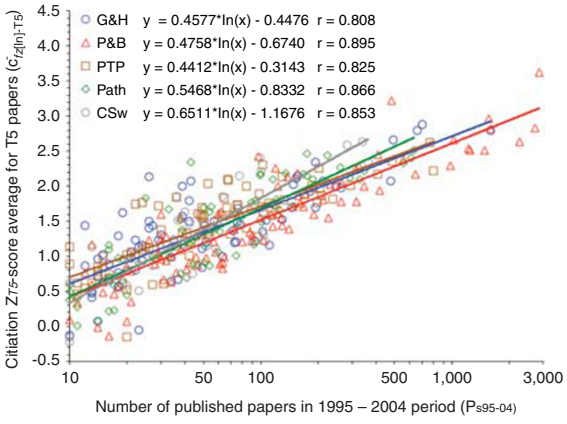


Figure 3 | Relationship between the number of published papers (P_{95-04}) and the citation z_{T5}-score averages ($\bar{c}_{fz[\ln]-T5}$) of five most highly cited papers.
***Notes:* Abbreviations for the research fields are the same as in Figure 1. All correlations are significant at $P < 0.0001$. The X-axis is in a logarithmic scale. Values of $\bar{c}_{fz[\ln]-T5}$ for species and research fields are in Supplementary Table S3.**

for all species, but species with a larger number of published papers would be expected to have a higher $\bar{c}_{fz[\ln]-T5}$ value. To remove variation in $\bar{c}_{fz[\ln]-T5}$ values caused by the number of published papers, 20,000 datasets were generated by randomly reshuffling citation data. These simulated datasets were then used to estimate citation z_{T5}-score averages ($\hat{c}_{fz[\ln]-T5}$), their confidence intervals and to test differences between simulated and observed citation z_{T5}-score averages. The simulated values of $\hat{c}_{fz[\ln]-T5}$ had similar profiles for all research fields (Figure 4), with somewhat slower growth observed for PTP and more rapid growth observed for CSw.

Numerical simulations of $\hat{c}_{fz[\ln]-T5}$ values yielded results highly consistent with the observed values of $\bar{c}_{fz[\ln]-T5}$ ($r = 0.841$, $P < 0.001$) (Figure 5). Significant differences between the observed and the simulated citation averages of *T*5 (shown as residuals; $r\bar{c}_{fz[\ln]-T5}$ in Supplementary Table S3) were observed in all research fields (Figure 5) and ranged from −1.25 for *Digitalis lanata* in PTP to 0.95 for *Malus domestica* in G&H. Positive values of $r\bar{c}_{fz[\ln]-T5}$ identify species cited more frequently than would be expected from simulated data, while negative $r\bar{c}_{fz[\ln]-T5}$ values indicate species cited less frequently. Ten species had $r\bar{c}_{fz[\ln]-T5}$ values significantly different from zero in two or more research fields (Supplementary Table S3). The values were always negative for five of these species (*Digitalis lanata, Hevea brasiliensis, Lens culinaris, Theobroma cacao* and *Trifolium repens*), one species had the values always positive (*Arabidopsis thaliana*) and four species showed mixed results (*Allium cepa, Humulus lupulus, Juglans regia* and *Prunus persica*). Species with significant $r\bar{c}_{fz[\ln]-T5}$ values have a substantial bias in citations of their most prominent papers, after adjusting for the number of published papers and distribution of citations in research fields. Such bias could be caused by many factors including, but not limited to a quality and type of research, preferential citation of species, publishing in less cited languages, publishing in non-indexed publications, differences in the time course of citations, relationship to extensively used species and growth rates in the number of publications.

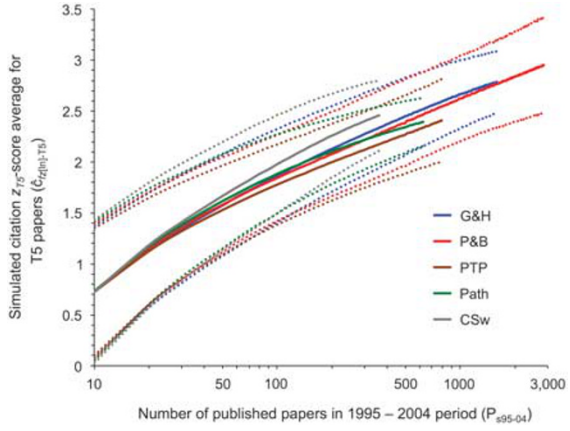


Figure 4 | Simulated citation z_{T5}-score averages ($\hat{c}_{fz[\ln]-T5}$) for five most highly cited papers.
***Notes*: Simulated averages (full lines) and their 95% confidence intervals (dotted lines) were determined from 20,000 datasets that have been generated by randomly reshuffling citation data within each research field. Abbreviations for the research fields are the same as in Figure 1. The *X*-axis is in a logarithmic scale.**

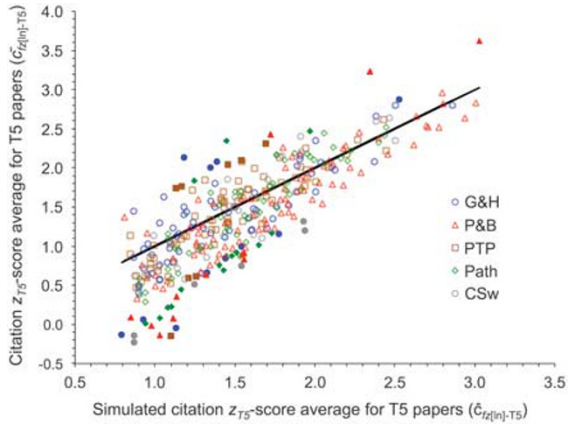


Figure 5 | Correlation between simulated and observed citation z_{T5}-score averages.
***Notes*: The black diagonal line shows values of a perfect match between simulated and observed data. Fully coloured symbols identify observed values that are significantly ($P < 0.05$) different from the simulated values (note that research fields have slightly different confidence intervals). Abbreviations for the research fields are the same as in Figure 1. The Pearson correlation coefficient between the simulated and the observed values is $r = 0.841$ ($P < 0.0001$). Differences between observed and simulated data (residual values, $r\bar{c}_{fz[\ln]-T5}$) are shown in Supplementary Table S3.**

RGR in share of publications. Analysis of RGR identified species with significantly changed share of publications from the earlier period (1995 to 2004) to the later period (2005 to 2013). The largest increase in the relative growth of publications was observed in G&H for *Populus trichocarpa* (RGR = 1.57), in P&B for *Arabidopsis lyrata* (1.32), in PTP for *Citrus sinensis* (0.88), in Path for *Lotus corniculatus* (0.53), and in CSw for *Solanum tuberosum* (0.95). Opposite, the largest decline in popularity was detected in G&H for *Brassica nigra* (RGR = −1.29), in P&B for *Taxus brevifolia* (−2.63), in PTP for *Petunia hybrida* (−1.70), in Path for *Quercus robur* (−1.63), and in CSw for *Lotus corniculatus* (−0.99). Forty-one species had a RGR significantly different from the overall mean in at least two research fields (Supplementary Table S3). Twenty-two of these species had significant RGR always negative (*Avena sativa, Betula pendula, Brassica oleracea, Daucus carota, Digitalis lanata, Helianthus annuus, Hordeum vulgare, Medicago sativa, Nicotiana tabacum, Petunia hybrida, Phaseolus vulgaris, Picea abies, Pisum sativum, Prunus amygdalus, Pyrus communis, Quercus robur, Secale cereale, Sinapis alba, Taxus brevifolia, Trifolium repens, Vicia faba*, and *Zea mays*), six species showed mixed results (*Beta vulgaris, Cucumis melo, Juglans regia, Lotus corniculatus, Solanum lycopersicum*, and *Solanum tuberosum*), while 13 species had significant RGR always positive (*Arabidopsis lyrata, Arabidopsis thaliana, Brassica rapa, Citrus sinensis, Gossypium hirsutum, Lotus japonicus, Malus domestica, Medicago truncatula, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Triticum aestivum*, and *Vitis vinifera*). Several of the species with consistently positive RGR are model organisms to study plant genetics, pathology, physiology, or biology, while a number of species with significant negative RGR were previously popular model organisms that are now used in research with lesser frequency.

Associations between bibliometric indicators. Tests of association were performed to identify relationships between the number of publications (P_{s95-04}), the RGR in the share of publications, and three citation indicators ($\bar{c}_{fz[\ln]}$, $\bar{c}_{fz[\ln]-T5}$ and $r\bar{c}_{fz[\ln]-T5}$). Association tests were performed on bibliometric indicators after classifying them into three groups based on their differences from overall means: significantly higher than the overall mean, non-

Table 2 | Associations between bibliometric indicators calculated from 108 species and 5research fields.

Indicator	$\bar{c}_{fz[ln]}$	$\bar{c}_{fz[ln]-T5}$	$r\bar{c}_{fz[ln]-T5}$	RGR
$P_{s95\text{-}04}$	0.09 ± 0.13	$0.95 \pm 0.02^{*}$	$0.39 \pm 0.14^{*}$	$-0.23 \pm 0.10^{*}$
$\bar{c}_{fz[ln]}$		$0.40 \pm 0.11^{*}$	$0.78 \pm 0.06^{*}$	$0.75 \pm 0.07^{*}$
$\bar{c}_{fz[ln]-T5}$			$0.89 \pm 0.03^{*}$	0.01 ± 0.09
$r\bar{c}_{fz[ln]-T5}$				$0.35 \pm 0.11^{*}$

Values of gamma (γ) ± standard errors are shown. * indicate associations significant at the experiment-wise $\alpha = 0.05$.
$P_{s95\text{-}04}$—number of published papers from 1995 to 2004, $\bar{c}_{fz[ln]}$—citation z-score average, $\bar{c}_{fz[ln]-T5}$—citation z_{T5}-score average for five most cited papers, $r\bar{c}_{fz[ln]-T5}$—residual of citation z_{T5}-score average for five most cited papers (difference between observed and simulated data), RGR—relative growth rate in the share of publications for a species.

significantly different from the overall mean and significantly lower than the overall mean, resulting in 3×3 contingency tables. Strong association ($\gamma = 0.95$, $P < 0.0001$) was detected between the number of publications ($P_{s95\text{–}04}$) and the citation z_{T5}-score average ($\bar{c}_{fz[ln]-T5}$; Table 2) confirming previous results of the Pearson correlation test (Figure 3). In addition a strong, positive association ($\gamma = 0.75$, $P < 0.0001$) was found between RGR in the share of publications (RGR) and citation z-score average ($\bar{c}_{fz[ln]}$). Trivial association was identified between $P_{s95\text{–}04}$ and $\bar{c}_{fz[ln]}$ ($\gamma = 0.09$), and between RGR and $\bar{c}_{fz[ln]-T5}$ ($\gamma = 0.01$). Weak, but significant negative association ($\gamma = -0.23$) was observed between the number of publications ($P_{s95\text{–}04}$) and the RGR indicating that species with a fewer publications have generally a faster growth rate. Notable exceptions to this trend are *Arabidopsis thaliana, Oryza sativa* and *Triticum aestivum* that each showed a significant positive RGR in at least two research fields, despite having a large number of publications in those fields. These data demonstrate that popularity of the three species is still growing regardless of their already frequent use in research. Contrary, 11 species showed a significant negative RGR in at least two research fields in which they have had a fewer papers than was the overall average for the field. These species are: *Avena sativa, Betula pendula, Digitalis lanata, Lotus corniculatus, Medicago sativa, Petunia hybrida, Prunus amygdalus, Quercus robur, Sinapis alba, Taxus brevifolia* and *Trifolium repens* (Supplementary Table S3). It appears that these species are used in research relatively less frequently, and interest in them is still declining.

The association observed in the present study, however, was not universal, and significant change in RGR was not always associated with a significant change in $\bar{c}_{fz[ln]}$ in the same direction. In two cases (*Nicotiana tabacum* in P&B and PTP) significant negative RGR was observed together with significantly high $\bar{c}_{fz[ln]}$, while in one case significant positive RGR was paired with significantly low $\bar{c}_{fz[ln]}$ (*Brassica rapa* in P&B). These discrepancies may reflect the real difference or they could be caused by inaccuracy of the two indicators (for example, RGR was computed from data within each research field only, while citations used in $\bar{c}_{fz[ln]}$ calculation could originate from all research fields). Similarly, there were 11 exceptions from the strong association between the number of published papers ($P_{s95\text{–}04}$) and the citation z_{T5}-score average for $T5$ papers ($\bar{c}_{fz[ln]-T5}$). In all exceptions a significantly fewer number of published papers was accompanied with significantly high $\bar{c}_{fz[ln]-T5}$ values (*Humulus lupulus* in PTP, *Lactuca sativa* in P&B, *Malus domestica* in G&H, *Medicago truncatula* in P&B and Path, *Nicotiana benthamiana* in P&B, *Petunia hybrida* in P&B, *Phaseolus vulgaris* in PTP, *Prunus persica* in G&H and *Vitis vinifera* in G&H and PTP).

Changes in popularity of experimental subjects. Because citation z-score average ($\bar{c}_{fz[ln]}$) was significantly associated with the RGR in the share of publications, additional analyses were performed to investigate changes in species share of publications from 1980 to 2013. Figure 6 illustrates examples of species from G&H research field with substantial increases in the share of publications (*Arabidopsis thaliana, Oryza sativa* and *Vitis vinifera*); substantial decreases in the share of publications (*Nicotiana tabacum, Petunia hybrida* and *Vicia faba*); or a relative constant share of publications (*Glycine max, Helianthus annuus* and *Sorghum bicolor*). In some genera, changes in the share of publications were highly different between species; stagnating or declining for some, while growing for those used as model organisms. For example, *Lotus corniculatus* share of publications declined, meanwhile *Lotus japonicus* share of publications increased (in large part due to studies of nitrogen fixation). A similar decline in share of publications was observed for *Medicago sativa*, while share of published papers with *Medicago truncatula*, a model organism, remarkably increased. Among trees, *Populus trichocarpa* become a popular species to study the genetics of woody plants, but the share of papers with *Populus deltoides* was steady (or declined if compared to the 1985–1989 period). Such substantial differences signify the importance of analysing the citations of species individually. In a preliminary study (data not shown) searches were performed for genus, instead of individual species. Genus searches have the advantage of getting more papers for statistical analyses from closely related species. Citation analyses of genus-based searches were similar to those performed on individual species, with the exception of genera where individual species had highly dissimilar RGR (for example, *Lotus*, or *Medicago*, Figure 6 and Supplementary Table S3). In genera where one of the species had high $\bar{c}_{fz[ln]}$ or RGR values, while the other species had low values of these two indicators, the $\bar{c}_{fz[ln]}$ and/or RGR values calculated on the genus level were similar to the overall mean of all species. In future, different species could show an increase or decrease in RGR. Such changes may occur due to the appearance of new model species (for example, *Brachypodium distachyon*) or a larger number of publications originating from countries preferring locally important plant species.

Discussion

The objective of the present study was to investigate the relationship between citations of published papers and plant species used in research. This was achieved by analysing a range of plant species from five research fields. Despite certain limitations, both the WoS and SCO databases provided a good approximation of the number of published papers and their citations. Similarly to earlier observations (Seglen, 1992; Redner, 2005; Lundberg, 2007), citations of published papers in all research fields showed lognormal distribution. Different aspects of citation distribution observed in other research fields (Hargens and Felmlee, 1984; MacRoberts and MacRoberts, 1989; Seglen, 1992; Redner, 1998; Adler *et al.*, 2009; Falagas *et al.*, 2010; Hsu and Huang, 2011) also apply to the present study. From a mathematical point of view, the citing of papers can be based on a relative simple model such as the model of cumulative advantage (Price, 1976) also known as the preferential attachment model (Barabási and Albert, 1999). In this model the probability that a paper is cited is proportional to the number of citations the paper has already received. However, the model does not explain certain details of citation distribution, thus the random-citing model has been proposed (Simkin and Roychowdhury, 2007). The random-citing model assumes that a researcher writing a manuscript cites several random recent papers and also copies some of their references. The model can also explain the citation pattern of papers that were cited little for a decade or more,

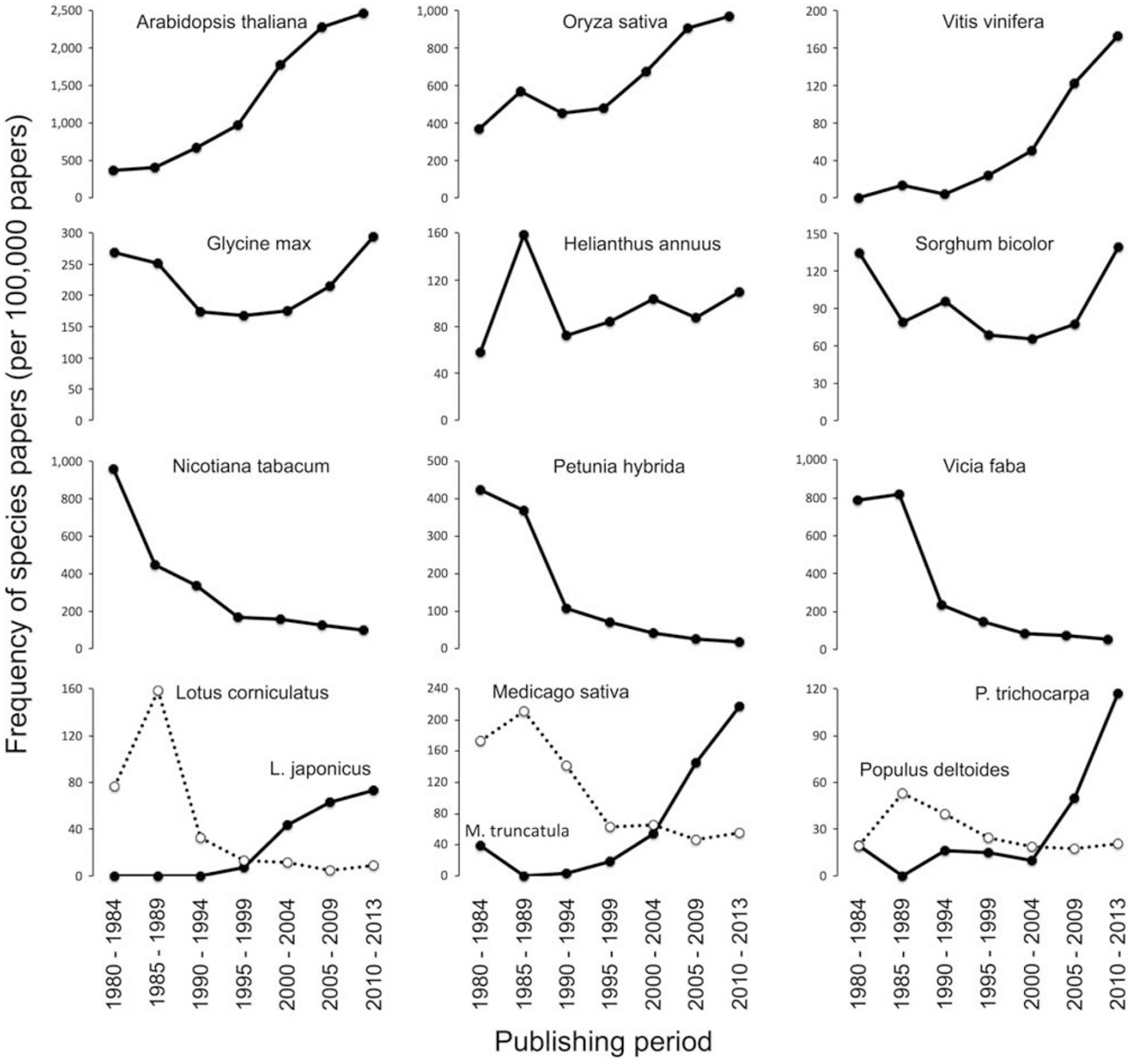
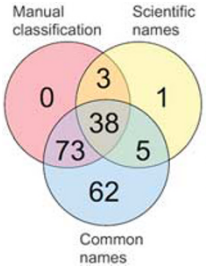


Figure 6 | Changes in species relative share of publications measured as frequency of published papers.
***Notes*: Analyses were performed for the G&H research field and the 1980 to July 2013 period. Share of publications is expressed as frequency of a species' papers per 100,000 G&H papers. Rows show examples of species with increasing popularity (top row), relative stable popularity (second row), decreasing popularity (third row). The bottom row shows genera where one species has increasing popularity while the other one has decreasing or relative stable popularity. Note the different scales for individual panels.**

but which get many recent citations (Simkin and Roychowdhury, 2007).

Significant differences among plant species were detected in both citation z-score average ($\overline{c}_{fz[\ln]}$) and citation z_{T5}-score average for $T5$ ($\overline{c}_{fz[\ln]-T5}$) in all research fields, including Crop Science journal with relatively narrow research field. Significant differences in citations of species were observed at five other journals that were analysed individually (HortScience, Plant Journal, Plant Physiology, Plant Science, and Theoretical and Applied Genetics; data not shown). The average number of citations of the five most prominent papers ($\overline{c}_{fz[\ln]-T5}$) of each species are strongly related to the number of published papers (P_{s95-04}); in other words, to the combined output of indexed papers produced by laboratories working with the species. This could lead to a perception that species frequently used in research also get the highest average number of citations. Analysis of citations, however, shows that the citation z-score average ($\overline{c}_{fz[\ln]}$) is not associated with the number of publications; rather it is significantly associated with the RGR in the share of publications (Table 2). Thus even species less frequently used in research, but with high RGR in the share of publications (often new model organisms) can reach very high $\overline{c}_{fz[\ln]}$ values. Conversely, previously popular model species with declining share of publications have usually low $\overline{c}_{fz[\ln]}$ values, even if citation of their $T5$ papers ($\overline{c}_{fz[\ln]-T5}$) is high.

The strong association between RGR and citation z-score average ($\overline{c}_{fz[\ln]}$) in this dataset is similar to observations in chemistry (Vinkler, 1996; Vinkler, 2002), stem cell research



Category	Scientific names	Common names
True positive	38 + 3 = 41	73 + 38 = 111
False positive	1 + 5 = 6	62 + 5 = 67
False negative	0 + 73 = 73	0 + 3 = 3

Figure 7 | Identification of species in databases using either plant common names or scientific names.
***Notes*: The Venn diagram shows a random sample of papers from the Web of Science (WoS) database that was categorized through an automatic classification or a manual classification performed by the author of this study. The automatic classification was carried out using either plant common names or scientific names in the field "Topic". The table shows the number of true positive, false positive, and false negative results of two automatic classifications assuming that the manual classification is correct. Note a large number of false positive results associated with the search that uses common plant names. Contrary, using plant scientific names leads to false negative results.**

(Barfoot *et al.*, 2013), physics, social sciences and archival fields (Hargens and Felmlee, 1984) where the relationship between growth in the number of publications and the mean number of citations per publication were reported. It was suggested that in rapidly expanding fields the average citation rate per paper is high because the number of citing papers is large relative to the amount of citable material (Hargens and Felmlee, 1984; Vinkler, 1996; Seglen, 1997). Similarity between previous observations at the research field level and current observations at the species level suggest that species should be considered as separate subfields within a research field when citations are analysed. This is a rather surprising finding considering that plant science publications often cite references from both the same (or closely related) species and from different genera. The same-species references are used to emphasize the importance of the species and to position current work in the context of previous achievements, while the cross-species references frequently include prominent papers from other species or genera. The high frequency of cross-species citations possibly play a role in significant $r\bar{c}_{fz[ln]-T5}$ values that indicate bias in citations of the most prominent papers after adjusting for the number of published papers and the distribution of citations in the research field. Prominent papers from the species with significant positive $r\bar{c}_{fz[ln]-T5}$ values are likely attracting frequent citations from other plant species. The difference in attracting citations from other species may also contribute to the fact that RGR was not always associated with a significant change in $\bar{c}_{fz[ln]}$ in the same direction. Comparable results were found at the research field level, where tendency of a research field to attract citations from adjacent fields was considered to be the most important factor affecting differences in citations among fields (Seglen, 1992; Seglen, 1997).

Because certain plant species may be preferentially used in basic research that is more likely to be cited than applied research (Narin *et al.*, 1976; Folly *et al.*, 1981; Vinkler, 1991), the ideal comparison of species' citations would include only papers describing identical research, published in the same journal at the same time, and preferably by the same authors. Such comparison is not possible, thus the present study has certain limitations that may be a factor in the observed distributions of citations and associations between indices: (a) only papers for specific time period were analysed; (b) the selection of papers was limited by database coverage and search criteria; (c) the grouping of journals into research fields was based on either classification provided by databases or personal experience of the author; (d) many crop-specific or industry-specific journals were not included in any of the research fields; (e) statistical analyses were performed on all identified papers regardless of the extent the species were used in research; (f) some papers may not be detected due to misspelled scientific names, change of scientific names or multiple synonyms of scientific names not identified by the author; and (g) the use of scientific names instead of common plant names underestimated the number of published papers (false negative results; Figure 7).

Conclusion

The six key findings of the current analyses of citations and bibliometric indicators are: (a) differences among citations of plant species from the same research field (or journal) are highly significant; (b) the greater the number of publications about a species, the higher the average citation score for the five most prominent papers of the species; (c) fast growth in the number of publications of a species leads to a high average citation score of the species; (d) plant model species (including *Arabidopsis thaliana*) have citation scores that are higher than the overall mean for the research field; (e) Funnel plot analysis is a convenient method to analyse citation data; and (f) computer simulations can approximate citations of the most prominent publications of a species.

The first four points are directly related to the key finding of the present work that shows that different species (particularly those with different volume and dynamics of publications) should be considered as separate subfields when analysing citation patterns within research fields. This observation likely applies to citation patterns in other areas of biological sciences. Therefore development of reliable bibliometric indicators that can compare performance of scientists working with different species is needed. Until such indicators are developed, committees allocating funding, and deciding appointments, promotions and tenures should take into consideration the fact that choice of experimental subject significantly influences citations of research papers.

References

Adler R, Ewing J and Taylor P (2009) Citation statistics: A report from the international mathematical union (IMU) in cooperation with the international council of industrial and applied mathematics (ICIAM) and the institute of mathematical statistics (IMS). *Statistical Science*; **24** (1): 1–14.

Althouse B M, West J D, Bergstrom C T and Bergstrom T (2009) Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*; **60** (1): 27–34.

Ball P (2007) Achievement index climbs the ranks. *Nature*; **448** (7155): 737–737.

Barabási A-L and Albert R (1999) Emergence of scaling in random networks. *Science*; **286** (5439): 509–512.

Barfoot J *et al* (2013) Stem cell research: Trends and perspectives on the evolving international landscape, http://www.elsevier.com/online-tools/research-intelligence/research-initiatives/stem-cell-research, accessed 24 April 2014.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*; **5** (1): 289–300.

Bornmann L, Mutz R and 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. *Journal of the American Society for Information Science and Technology*; **59** (5): 830–837.

Campbell P (2008) Escape from the impact factor. *Ethics in Science and Environmental Politics*; **8** (1): 5–6.

Cole S, Cole J R and Simon G A (1981) Chance and consensus in peer review. *Science*; **214** (4523): 881–886.

Falagas M, Kouranos V, Michalopoulos A, Rodopoulou S, Batsiou M and Karageorgopoulos D (2010) Comparison of the distribution of citations received by articles published in high, moderate, and low impact factor journals in clinical medicine. *Internal Medicine Journal*; **40** (8): 587–591.

Folly G, Hajtman B, Nagy J and Ruff I (1981) Some methodological problems in ranking scientists by citation analysis. *Scientometrics*; **3** (2): 135–147.
Garfield E (1972) Citation analysis as a tool in journal evaluation. *Science*; **178** (4060): 471–479.
Glänzel W (1996) A bibliometric approach to social sciences. National research performences in 6 selected social science areas, 1990-1992. *Scientometrics*; **35** (3): 291–307.
Hargens L L and Felmlee D H (1984) Structural determinants of stratification in science. *American Sociological Review*; **49** (5): 685–697.
Hirsch J E (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*; **102** (46): 16569–16572.
Hsu J-w and Huang D-w (2011) Dynamics of citation distribution. *Computer Physics Communications*; **182** (1): 185–187.
Iglesias J E and Pecharromán C (2007) Scaling the *h*-index for different scientific ISI fields. *Scientometrics*; **73** (3): 303–320.
Jin B, Liang L, Rousseau R and Egghe L (2007) The R-and AR-indices: Complementing the *h*-index. *Chinese Science Bulletin*; **52** (6): 855–863.
Lundberg J (2007) Lifting the crown—citation *z*-score. *Journal of Informetrics*; **1** (2): 145–154.
MacRoberts M H and MacRoberts B R (1989) Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*; **40** (5): 342–349.
Narin F, Pinski G and Gee H H (1976) Structure of the biomedical literature. *Journal of the American Society for Information Science*; **27** (1): 25–45.
Neff M W and Corley E A (2009) 35 years and 160,000 articles: A bibliometric exploration of the evolution of ecology. *Scientometrics*; **80** (3): 657–682.
Price D d S (1976) A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*; **27** (5): 292–306.
Radicchi F, Fortunato S and Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*; **105** (45): 17268–17272.
Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*; **4** (2): 131–134.
Redner S (2005) Citation statistics from 110 years of Physical Review. *Physics Today*; **58** (6): 49–54.
Reed K L (1995) Citation analysis of faculty publication: beyond Science Citation Index and Social Science Citation Index. *Bulletin of the Medical Library Association*; **83** (4): 503–508.
Rehn C, Kronman U and Wadskog D (2007) *Bibliometric indicators—definitions and usage at Karolinska Institutet*; Karolinska Institutet University Library: Stockholm, Sweden.
Reinstein A, Hasselback J R, Riley M E and Sinason D H (2011) Pitfalls of using citation indices for making academic accounting promotion, tenure, teaching load, and merit pay decisions. *Issues in Accounting Education*; **26** (1): 99–131.
Schubert A, Glänzel W and Braun T (1989) Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. *Scientometrics*; **16** (1): 3–478.
Seglen P O (1992) The skewness of science. *Journal of the American Society for Information Science*; **43** (9): 628–638.
Seglen P O (1997) Citations and journal impact factors: questionable indicators of research quality. *Allergy*; **52** (11): 1050–1056.
Simkin M V and Roychowdhury V P (2007) A mathematical theory of citing. *Journal of the American Society for Information Science and Technology*; **58** (11): 1661–1673.
Simko I (2015) Citations of 108 plant species in five research fields, Harvard Dataverse, V1. http://dx.doi.org/10.7910/DVN/GHOHJI.
Spiegelhalter D J (2005) Funnel plots for comparing institutional performance. *Statistics in Medicine*; **24** (8): 1185–1202.
Stringer M J, Sales-Pardo M and Amaral L A N (2008) Effectiveness of journal ranking schemes as a tool for locating information. *PLoS One*; **3** (2): e1683.
Tijssen R J, Visser M S and van Leeuwen T N (2002) Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*; **54** (3): 381–397.
Vinkler P (1988) Bibliometric features of some scientific subfields and the scientometric consequences therefrom. *Scientometrics*; **14** (5): 453–474.
Vinkler P (1991) Possible causes of differences in information impact of journals from different subfields. *Scientometrics*; **20** (1): 145–161.
Vinkler P (1996) Relationships between the rate of scientific development and citations. The chance for citedness model. *Scientometrics*; **35** (3): 375–386.
Vinkler P (2002) Dynamic changes in the chance for citedness. *Scientometrics*; **54** (3): 421–434.

Data Availability

The datasets analysed during the current study are available in the Harvard Dataverse repository (Simko I, 2015): http://dx.doi.org/10.7910/DVN/GHOHJI. All data generated during this study are included in this published article or in the supplementary information.

Acknowledgements

Egan McComb wrote the Python code and Adam Šimko made corrections to the manuscript. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Additional information

Supplementary Information: accompanies this article at http://www.palgrave-journals.com/palcomms.

Competing interests: The authors declares no competing financial interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Ivan S (2015) Analysis of bibliometric indicators to determine citation bias. *Palgrave Communications* 1:15011 doi: 10.1057/palcomms.2015.11.