NOAA Technical Memorandum OAR PPE-4

Bibliometric Performance Measures for the Evaluation of NOAA R&D

C. Belter and A. Sen

July, 2014

FOR MORE INFORMATION:
For more information about this report or to request a copy, please contact NOAA's Office of Oceanic and Atmospheric Research, Office of Policy, Planning, and Evaluation: NOAA; R/PPE; 1315 East West Highway; Silver Spring, MD 20910 or visit www.research.noaa.gov.

DISCLAIMER:
Mention of trade names or commercial products does not constitute endorsement or recommendation for their use by the United States government.

This document is available to the public through:
The NOAA Deepwater Horizon Institutional Repository
http://noaa.ntis.gov

U.S. Department of Commerce
Penny Pritzker, Secretary
National Oceanic and Atmospheric Administration
Kathryn Sullivan, Under Secretary and Administrator
Oceanic and Atmospheric Research
Craig McLean, Acting Assistant Administrator

Bibliometric Performance Measures for the Evaluation of NOAA R&D

Christopher Belter
Avery Sen

Abstract

The purpose of this paper is to recommend the adoption of a suite of bibliometric indicators to assist in the evaluation of NOAA R&D. We select a suite of indicators based on the recommendations and best practices established in the bibliometric literature and provide a framework for implementing these metrics in evaluation procedures. A suite of metrics for the evaluation of NOAA's R&D enterprise can be organized into four overarching themes: production, collaboration, topicality, and quality. We suggest that these indicators be adopted at multiple levels of the agency to not only evaluate the entire agency's publication output, but also that of NOAA's line offices and individual R&D units. This could lead to a more systematic adoption and use of bibliometric indicators beyond NOAA, at other US government agencies.

Introduction

NOAA is a science-based federal mission agency. Research and development (R&D) underlie its diverse responsibilities to predict the weather, regulate fisheries, assess climate impacts, protect coastal ecosystems, and operate environmental satellites. There are many complexities inherent in the science of oceans and atmosphere, and these complexities are reflected, and in many ways magnified by in the interrelated functions and organizations that constitute NOAA's R&D enterprise.

NOAA has over a hundred different mandates and authorities, but no single legislative "organic act" to define its purpose and responsibilities comprehensively. The agency was created by executive order to integrate a number of pre-existing bureaus, several of which, such as the Weather Bureau and the Coast Survey, have pedigrees extending back nearly as far as the Nation itself. Different operational lines have unique cultures, established procedures, and varied talent. While this diversity can be a source of organizational strength and resilience, it can also result in intra-organizational friction and high transaction costs. This is especially true of NOAA's R&D endeavors, which are distributed across five line offices, and which vary in applicability to operational service improvements.

Evaluation of R&D at NOAA is more than the practice of "good government," it is a medium for the agency's continued evolution as a single entity -- a whole that is more than the sum of its parts. Through shared ambitions, joint analysis of interdependent value-chains, as well as transparent, evidence-based decision-making, the practice of evaluation has the potential to unify an agency in which different cultures and competing interests are at play.

The purpose of this paper is to recommend the adoption of a suite of bibliometric indicators to assist in

the evaluation of NOAA R&D. Not only do we select a suite of indicators based on the recommendations and best practices established in the bibliometric literature, we also provide a framework for implementing these metrics in evaluation procedures. In the process, we also hope to provide guidelines for the adoption of bibliometric indicators at other government agencies. Although specific evaluation procedures ought to be tailored to the unique circumstances surrounding the R&D enterprise at each individual agency, the evaluative framework and, to some degree, the indicators selected here could be used to frame the evaluation process at other agencies. This could lead to a more systematic adoption and use of bibliometric indicators across US government agencies.

The need to create a culture of evaluation at NOAA cannot be separated from the larger federal endeavor to do so. The past few years have seen Congress and the President reassert the importance of performance management and evidence-based decision-making in government. The Government Performance and Results Modernization Act (2010) placed new conditions on strategic planning, program performance and evaluation for agencies. The Office of Management and Budget (OMB) Circular A-11 (2013) requires agencies to maintain a decision-making process that integrates analysis, planning, evaluation, and budgeting. OMB Memorandum M-10-01 (2009) identified evaluation as a keystone to "determine how to spend taxpayer dollars effectively and efficiently—investing more in what works and less in what does not." OMB Memorandum M-10-24 (2010) encouraged agencies to use performance information to lead, learn, and improve outcomes; communicate performance coherently and concisely for better results and transparency; and strengthen problem-solving networks, inside and outside government, to improve outcomes and performance management practices. The National Research Council (2012) points to a number of specific, system-wide criteria with which to assess the management, quality, and impact of research endeavors.

Despite frequent recommendations from the academic community, and the formation of the STAR Metrics program (Lane 2010; Lane 2011), the adoption of bibliometric indicators for evaluating R&D conducted and supported by agencies of the US government has been sporadic. In most cases, scientometric indicators have either been used on an *ad hoc* basis by certain offices within US federal agencies, or calculated by academics on behalf of these agencies or subagencies. Unsurprisingly, the majority of this work has been conducted on research supported by NIH (e.g. Boyack 2003; Boyack 2011; Druss 2005; Herr, 2009; Liebow 2009; Lyubarova 2009; Rosas 2011; Yang 2013) and NSF (e.g. Huang 2005; Huang 2006; Youtie 2013; Zoss 2012), but analyses have also been conducted at USDA (Kosecki 2011) and NOAA's Office of Ocean Exploration and Research (Belter 2013). Bibliometric analyses and evaluations have also been conducted for departments within NIST and FDA, but we are not aware of publications resulting from these efforts.

We emphasize that we are not recommending that bibliometric indicators be adopted as a replacement for peer review in the evaluation of NOAA R&D. Due to the well-established limitations of bibliometric indicators (e.g. Adler and Harzing 2009; Leydesdorff 2008; van Raan 2005a), the bibliometric literature recommends that these indicators be used in combination with peer review to evaluate scientific research (e.g. Derrick 2013; Haeffner-Cavaillon 2009; Moed 2007; van Raan 1996). In addition to correcting for the limitations of bibliometrics, the combination of bibliometrics and peer review helps correct for the resource intensiveness, limited scope, and potential biases of peer review (e.g. Bornmann 2011b; Lee 2013), as well as foster the democratization, openness, and replicability of scientific research evaluations

(Derrick 2013). Although our focus here is on utilizing bibliometric indicators for research evaluation, we recommend that these metrics be implemented alongside NOAA's existing peer review system to gain a more comprehensive perspective on, and provide a more holistic assessment of, NOAA's R&D enterprise.

A Suite of Measures

It has been well established that using bibliometric indicators to evaluate scientific research has a wide range of intended and unintended effects on the subsequent publishing behavior of the scientists being evaluated (e.g. Bornmann 2011a; Butler 2003; Jimenez-Contreras 2002; Kostoff 2007; Moed 2008; van Dalen 2012; Weingart 2005). Implementation of productivity indicators, such as publication counts, without accompanying quality indicators resulted in researchers publishing more articles in lower-impact journals and splitting the results of their research into multiple publications that might otherwise have been published as a single article. In response to incentives to publish in journals with high impact factors - which, in our opinion, is a deeply flawed approach (e.g. Opthof 1997; Seglen 1997) - researchers have tended to perform short-term, conventional research, as opposed to pursuing more controversial, long-term, or speculative lines of research, to increase the probability of being accepted by such journals.

A suite of indicators measuring a broad range of publication characteristics implemented in combination with qualitative evaluation techniques, such as peer review, may be able to avoid such unintended effects. The bibliometric literature recommends implementing a broad suite of metrics for a more holistic understanding of the articles produced by an author or institution because each bibliometric indicator measures a different aspect of the underlying publication set (e.g. Martin, 1996; van Leeuwen 2003). A suite of metrics, then, can provide indications of, and credit for, a number of these various aspects, hopefully encouraging researchers to produce articles that contribute to multiple aspects.

A suite of metrics for the evaluation of NOAA's R&D enterprise can be organized into four overarching themes: production, collaboration, topicality, and quality. The production indicators attempt to measure the amount of research performed over specified periods of time. The collaboration indicators attempt to identify both the amount of knowledge shared with researchers outside of NOAA and the major institutional, sectoral, and international partners with whom this knowledge is shared. The topical indicators attempt to identify the major research areas pursued by NOAA R&D and the distribution of research effort across these research areas to ensure that they are consistent with the research priorities set forth in NOAA's Next Generation Strategic Plan. Finally, the quality indicators attempt to measure the value of NOAA R&D to the broader scientific community. [1]

We suggest that these indicators be adopted at multiple levels of the agency to not only evaluate the entire agency's publication output, but also that of NOAA's line offices and individual R&D units. The actual

[1] It is important to note that "quality" is often a subjective assessment. As will become clear, to be more objective, we understand quality as the utility of research to other research, as evidenced by citations. We will never have "perfect" knowledge of the quality of a research publication, but we can have less imperfect knowledge. Citations are an imperfect but informative indicator of quality. There may also be non-bibliometric ways of getting even less imperfect knowledge of quality, but that is not the subject of the paper.

metrics we recommend are as follows:

Productivity: How many articles are published?

- Number of publications per unit of time
- Number of publications per unit of aggregation (R&D unit, LO, agency)

Collaboration: With whom are articles published?

- Percentage of publications co-authored across R&D units (intramural)
- Percentage of publications co-authored across line offices (intramural)
- Percentage of publications co-authored across agencies/institutions (extramural)
- Percentage of publications co-authored across sectors (extramural)
- Percentage of publications co-authored across countries (extramural)

Topics: What are the articles about and how do they integrate disciplines?

- Number of publications per subject area (predefined)
- Number of publications per research topic (emergent)
- Number of subject areas citing publications
- Number of subject areas cited by publications
- Ratio of citations within vs. outside of publications' subject area

Quality: How good are the articles, per the scientific community?

- Number of citations
- Number of citations per publication
- Total percentage of publications in the top ten percent
- Percentage of papers in the top ten percent, per subject area (predefined)

In the following sections, we describe each of these measures, lay out our rationale for selecting them, and provide an overview of their limitations. We then discuss how these measures can be applied to assess NOAA's R&D enterprise.

Productivity

Number of Publications

Publication in peer reviewed journals, books, and other venues (hereafter referred to simply as "publication") is an indicator of legitimacy. Publication indicates that the work produced by an author or R&D unit is both significant and scientifically rigorous enough to warrant distribution to the scientific community. Publication counts over specified periods of time are therefore indicators of both the productivity and the intellectual contribution of authors or groups. Publication counts can be generated for various levels of NOAA's organizational structure - authors, R&D units, line offices, and the entire agency - and can provide an indication of the balance of NOAA's intramural and extramural research.

It is important to note that publication counts are dependent on the database used to generate

them. Differences in the coverage and the metadata quality among databases can lead to substantial differences in the number of publications retrieved from different databases. Although we recommend using Web of Science as the primary data source for bibliometric indicators at NOAA because of its availability at NOAA, breadth of coverage, metadata quality, and citation analysis capabilities, our subscription to Web of Science does introduce important limitations on the scope of those indicators. Our subscription to Web of Science does not include coverage of book chapters or publications in the social sciences. In addition, the coverage of Web of Science, though broad, does not include all of the scientific journals in which NOAA research is published and is known to be biased toward English-language publications (Liang 2013; van Leeuwen 2001; van Raan 2011). Because of these limitations, metrics calculated using Web of Science data represent an incomplete subset of NOAA's actual publication output.

Collaboration

Number of publications co-authored across R&D units, line offices, agencies, sectors

Co-authorship on scientific publications is often used as an indicator of scientific collaboration at the intra-organizational, inter-organizational, and international scales (e.g. Schubert 1990; Glanzel 2001; Melin 1996; Mahlck 2000; Wagner 2005; The Royal Society 2011; Chen 2013). Co-authorship on scientific publications indicates the sharing or shared creation of knowledge and/or expertise between individuals, and, by extension, between the individuals' institutions and countries. Although certain highly-productive individuals can have profound effects on co-authorship patterns at all scales of aggregation (Chen 2013), co-authorship can nonetheless identify the major trends and partnerships at each these scales. Co-authorship also has the advantage of creating networks that can be analyzed and visualized using methods developed by the broader field of network science (Albert 2002; Borner 2007; Newman 2003).

We recommend that co-authorship metrics and networks be generated at five scales: intramural research unit (e.g. SWFSC, GFDL, etc.), intramural line office (e.g. NOS, NESDIS, etc), organization (e.g. NASA, University of Colorado, etc), sector (government, academic, or private), and country. Co-authorship metrics and networks at each of these scales would allow us to determine the degree to which NOAA R&D units and line offices are already co-publishing, ascertain their major co-authorship partners, and identify potential partners for future research publication at each of these scales. The actual metrics we suggest are the absolute number of co-authored publications, the percentage of co-authored publications as compared to all publications, and the identification of the institutions, sectors, and countries with which NOAA most often co-publishes. We also recommend using network analysis and visualization techniques to identify the structure of these partnerships.

The limitations of using co-authored publications as an indicator of collaboration are primarily conceptual (Melin 1996). Collaboration is a [complex] concept that can be manifested in many ways. Co-authorship on published articles is only one form of collaboration and too much emphasis on it could potentially obscure other, more organic, forms of collaboration. At the same time, co-authorship is not necessarily an indicator of direct collaboration. Articles by tens to

hundreds of authors are becoming increasingly common in scientific publishing (citation). Individual authors on such articles may not ever directly interact with one another, although they can in some sense be called collaborators since they contributed to a common product.

Finally, all indicators of co-authorship are prone to errors related to the inconsistencies and ambiguities of author names and addresses. Articles may be erroneously attributed to authors with the same names (Wang, Z. or Smith, P) and articles by the same author may be attributed to different authors due to inconsistencies or misspellings of author names (Belter, CW vs. Belter, C). In addition, single authors may have multiple affiliations, resulting in institutional 'collaboration' on single-authored articles. Although the effects of such errors on co-authorship indicators are difficult to determine, it is likely that such indicators are at least broadly accurate.

Topics

Number of publications, per subject area (predefined)

Each individual publication in Web of Science is automatically assigned to broad subject area(s) based on the journal or book series in which it is published. Examples of subject areas in which NOAA authors routinely publish include 'Meteorology and Atmospheric Sciences', 'Marine and Freshwater Biology', and 'Oceanography.' Depending on the journal in which it is published, an article may be assigned to a single or multiple subject area, but all articles published in the same journal will be assigned to the same subject area(s). For example, an article published in Marine Ecology Progress Series will be assigned to the 'Ecology', 'Marine and Freshwater Biology', and 'Oceanography' subject areas, whereas an article published in the Journal of Experimental Marine Biology and Ecology will be assigned to the 'Ecology' and 'Marine and Freshwater Biology' subject areas.

Counts of publications in each of these subject areas can give a broad indication of the overall distribution of NOAA publications across disciplines. While useful, this metric has several limitations. The subject areas defined by Web of Science may be too broad to provide a clear indication of the actual topics pursued by NOAA publications. Because article classifications are defined based on the journal in which the articles are published, there is a degree of error involved: an article published in Marine Ecology Progress Series may not be related to Oceanography, but it will be assigned to that subject area because of the journal in which it was published. Finally, there is some evidence that the subject categories defined by Web of Science may not necessarily match those that emerge from the analysis of citation networks (Boyack 2005).

Number of publications, per research topic (emergent)

In an attempt to correct for these limitations, we also propose that NOAA articles be classified according to topics that emerge at the article level. Classification of articles by journal presents two of the same issues as doing so by subject area - lack of sufficient topic granularity and the potential for erroneous classification - and cannot efficiently classify articles published in

multidisciplinary journals such as Science or Geophysical Research Letters. Instead, we propose that these articles be assigned to topics using the bibliometric mapping (Borner, 2003) technique called bibliographic coupling (Kessler, 1963). Bibliographic coupling leverages the fact that related papers typically cite the same previous literature, allowing the topical structure of a set of documents to emerge naturally from within the set. In bibliographic coupling, if papers A and B both cite paper C, a link is created between A and B. The more papers A and B both cite, the stronger the link between A and B becomes, and the more likely it is that they cover the same, or similar, topics. We recommend bibliographic coupling because of its accuracy in identifying topical structure (Boyack, 2010) and its ability to analyze recently published articles. The resulting bibliographic coupling network can then be analyzed using a community detection algorithm to identify its major research topics and visualized to depict the topical structure and the distribution of publications across topics.

The topics identified using this method can then be coded to one or more of NOAA's strategic goals (climate, weather, oceans, and coasts) or to it cross-cutting objectives to evaluate the balance of NOAA's publication activity per goal. To do so, we would manually associate each topic, and the publications on each of these topics, with one or more goals and then calculate the number of publications contributing to each goal. The final tallies can then be visualized as a network to show both the number of publications associated with each goal and the number of publications that cross-cut these goals, allowing us to see the number of publications that contribute, for example, to both NOAA's weather and climate goals.

Number of subject areas citing the publications (or cited by it)

NOAA's objective for "a holistic understanding of the earth system" presents a unique challenge for performance evaluation. How do we know if understanding of the earth system is "holistic?" The word is employed from the perspective of system theory to distinguish the properties of wholes from those of the sum of their parts (Von Bertalanffy, 1950). Thus, a holistic understanding means something different from a *precise* understanding (i.e., increasing the fidelity with which we understand phenomena) and from a *fundamental* understanding (i.e., reducing phenomena to general principles). The objective is also different from a *comprehensive* understanding, that is, understanding all aspects of all elements of a system. Striving for holistic understanding is not the same as striving for a more complete understanding (i.e., omniscience).

Rather, it means an understanding of the Earth system that is an integration of other, often pre-existing and always incomplete, forms of understanding. The research to produce holistic understanding is synthetic, in addition to (and instead of) research that is analytic -- *connecting* the dots, not *collecting* the dots. For ecosystems, it is the study of ecology as distinct from species or habitats. For climate systems it is the bridging of phenomena at different spatial and temporal scales. For observation systems, it is the architecture of interwoven technologies to collect and manipulate data. Metrics for holistic understanding must, therefore, account for cross-disciplinary integration, or, as it is often referred to in the bibliometrics literature, interdisciplinarity. For a review of bibliometric methods of measuring interdisciplinarity, see Wagner (2011).

The first proposed metric is the breadth of different subject areas that cite a particular publication. It indicates how relevant the publication, or group of publications, is across professional communities, and thus the diversity of potential influence of the underlying research. It is calculated by first establishing the set of publications that have cited the work(s) in question, then counting the number of subject areas represented in this set. This metric is not intended to indicate the depth of influence in any particular subject area because, as noted above, different domains have different rates of publication. Rather, it is intended to indicate the span of influence that the research might have.

The second, corresponding measure of the breadth of different subject areas that are cited by a particular publication. While the first measure above is one of "outbound" diversity, this measure is one of "inbound" diversity, that is, the diversity of knowledge underlying the research published. It indicates the extent to which research is based on an integrated understanding of a number of subject areas.

Ratio of citations within vs. outside of publications' subject area

Also in keeping with NOAA's desire to measure a holistic understanding of the Earth system, as well as the assumption that this entails accounting for cross-disciplinary integration, this metric focuses on the proportion of "home field" citations to those of "away fields." More specifically, the metric is calculated as the number of citations of an article (or set of articles) that emerge from the subject area of each article divided by the number of citations that emerge from different subject areas. Because of the different rates of publication across disciplines, this metric is defined as a ratio, rather than an absolute number. Further, this metric does not specify among disciplines; it only accounts for "in" versus "out."

As with the previous metric, this one can also have a corresponding "inbound" metric, that is, the ratio of citations contained within the article (or set of articles) being analyzed. It too, could be used to indicate the extent to which research is based on an integrated understanding of a number of subject areas.

Although the bibliometric research community has yet to come to a consensus on the best method for measuring the interdisciplinarity of a document set, the four metrics we have proposed to do so for NOAA publications are the four metrics most commonly used for this purpose in the bibliometric literature (Wagner, 2011).

Quality

Number of Citations

The citation of one publication by another is an indication that the publication being cited is both of sufficient quality to deserve mention and in some way exerted some influence on the publication citing it. The total count of citations received by a publication, therefore, provides an

indication of the quality of that publication as estimated by the scientific community. The total count of citations received by all articles in a set of publications - such as that by an author, institution, or country - provides an indication of the quality of the entire set and, by extension, the author or institution to whom the set belongs.

Although authors may cite publications for reasons having nothing to do with acknowledging intellectual influence (e.g. Bornmann and Daniel 2008), van Raan (2005) argues that citation counts are still a valid indicator of quality, in part because of the broad agreement between citation metrics and peer judgement (e.g. Bornmann 2011; van Raan 2006). Still, citation counts, and all of the metrics derived from citation counts, can only provide an indication of the quality of scientific publications due to the uncertainty surrounding citation motivation(s). As a result, citation counts and citation-derived metrics should be interpreted with caution.

In addition, raw citation counts are influenced by a number of factors beyond publication quality. First, citation counts are dependant on discipline. Publications in some disciplines tend to receive more citations, on average, than those in other disciplines. This is because differences in the number of publications produced and the length of average cited references lists among disciplines translate to higher numbers of available citations - or, a higher 'citation potential' (Garfield, 1979) - in some disciplines than in others. The practical implication of these differences is that an article with 10 citations in Chemistry might have an average number of citations, but an article with 10 citations in Mathematics might be highly cited. As a result, raw citation counts, and citation indicators that are not normalized to account for such differences, cannot be compared across disciplines.

Second, citation counts are dependant on time. Citation counts increase over time as new articles are published and can never decrease. This means that older articles are, on average, more highly cited than newer ones. An article in Chemistry with 10 citations might be highly cited if it had been published last year, but may have an average number of citations if it had been published 3 years ago. See Figure 1 for an illustration of the effects of age and discipline on the median citation count of articles published in the five disciplines in which NOAA authors most frequently publish. In addition, publications typically require between 2 and 5 years to reach their peak citation rates (Costas, 2011; Eom and Fortunato, 2011). This means that citation counts, and bibliometric indicators based on citation counts, cannot be considered stable enough to be used for evaluative purposes until at least 2 years after an article's publication date. Most methodologically rigorous bibliometric evaluations analyze publications produced over a 5 or 10 year span in which the most recent publications are at least 2 years old.

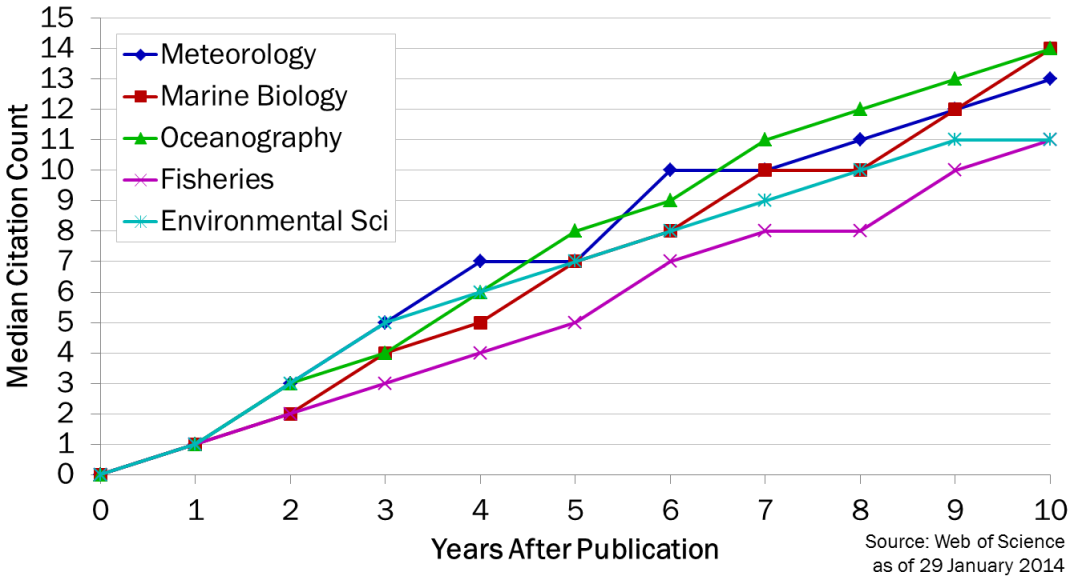

Figure 1: Median citation counts for articles published in five scientific disciplines from 2003 to 2013.

Third, citation counts are dependant on publication-related factors. Certain types of publications tend to receive more citations than others: review articles tend to be highly cited, publications reporting research results tend to receive an average number of citations, and other types of publications (letters, editorials, notes, etc) tend to receive fewer citations. In addition, citation counts generated for sets of publications tend to be dependent on the number of publications in the set. A large set of publications has more opportunities to be cited, and will therefore tend to receive a larger number of citations, than a comparable set of fewer publications. For that reason, citation counts for sets of publications ought to be accompanied by additional bibliometric indicators that correct for the number of publications in those sets.

Citations per Publication (CPP)

One popular and well-established method of correcting for the size of a publication set is to calculate the average number of citations each publication in the set has received. This is done by simply dividing the total citation count of all of the publications in the set by the number of publications in the set. This metric is used to summarize the collective impact of a set of publications on subsequent research. By accounting for both the number of publications in and the number of citations received by a set of publications with a single number, CPP attempts to correct for the fact that larger publication sets tend to receive a greater number of citations than smaller ones.

Although frequently used, this metric also has limitations. Because it is based on raw citation counts, CPP cannot be compared across disciplines. In addition, because the distribution of citations among articles is highly skewed (e.g. Albarrán and Ruiz-Castillo 2011; Redner 1998; Seglen 1992), this and other indicators based on averages may give a distorted indication of the actual distribution of citations among a set of papers. CPP is easily influenced by a few very

highly cited papers, or a large number of seldom cited papers. Finally, citation counts and CPP by themselves are not meaningful because they provide no context; they only become meaningful when compared to indicators calculated for a reference set of similar publications.

Percentage of Publications in the Top Ten Percent ($PP_{Top10\%}$)

The need to correct for differences in the citation potential among disciplines has resulted in a sizable literature on methods of normalizing citation counts in various ways. Traditionally, such normalization was accomplished by dividing the CPP of a collection of papers by the CPP of either the journals in which these publications appeared or the subject categories to which these papers were assigned (Moed 1995). Although this method was recently updated (Waltman 2011a, Waltman 2011b), it still relies on the calculation of averages, which, as noted previously, are highly sensitive to the skewed distribution of citations among publications.

In the recent debate surrounding this method, a new method of normalizing citation counts by means of percentile ranks was proposed (Leydesdorff 2011). Inspired by the methods used in the biannual Science and Engineering Indicators report published by the National Science Board (2012), among other sources, this new method measures the percentage of publications in a particular set of publications that have citation counts ranking in a predetermined percentile, or set of percentiles, as compared to all publications in a larger set of publications. In practice, this method has quickly evolved into measuring the percentage of publications by a particular research group, institution, or country with citation counts ranking in the top 10% of all publications of the same publication type, year of publication, and subject area. This indicator has been adopted by the Center for Science and Technology Studies (CWTS)' Leiden Ranking as the $PP_{Top10\%}$ indicator (Waltman 2012) and the SCImago Institutions Rankings as the Excellence Rate (Bornmann 2012). It is considered by both groups to be the most stable and accurate bibliometric quality indicator currently available. By measuring the percentage of publications with citation counts above a constant percentile threshold, this indicator theoretically corrects for most of the known limitations of citation counts and allows for straightforward interpretation of the resulting indicator: below 8% indicates below average performance, 8-12% indicates average performance, and above 12% indicates high performance.

Although this is likely to be the best bibliometric indicator of quality currently available, it does have limitations. Because multiple publications often have the same citation count, the percentile threshold for the top 10% is not always clearly delineated. That is, the percentage of all papers with citation counts at or above the 10% threshold may, and often does, exceed 10%. Although several methods of correcting for this issue have been proposed, it is too soon to know which will eventually become accepted practice. In addition, this indicator may not sufficiently control for differences in citation distributions among fields, as there is some evidence that high rates of publication in certain subject areas may result in an advantage in an institution's overall ranking (Bornmann, 2013). Finally, since the suggestion and implementation of this indicator is so recent, it is probable that additional issues will arise over time.

Percentage of papers in the top ten percent, per subject area (predefined)

In addition to calculating the $PP_{Top10\%}$ indicator for all publications in a given set, we also suggest calculating this percentage for each of the major subject categories to which those articles have been assigned by WoS. This more granular approach allows for the evaluation of NOAA publications by subject area to identify the subject-specific strengths and weaknesses of NOAA publications. Subject-specific percentile ranks have the same advantages and limitations identified above for general percentile ranks.

Conclusion

The suite of metrics recommended here can be organized according to the intellectual framework provided by a recent report by the National Research Council (2012). The report identifies three overarching elements for the assessment of R&D organizations: management, quality, and impact, which largely parallel NOAA's assessment criteria of performance, quality, and relevance. Assessing the management of an organization involves ensuring that the work performed by the organization supports the organization's strategic goals and that the organization has the sufficient workforce, infrastructure, and leadership resources necessary to achieve those goals. Assessing the quality of the work done involves evaluating the value of the organization's outputs and by benchmarking the organization against other organizations of similar size and scope. Assessing the impact of an organization involves measuring the broader societal benefits of that organization's work.

It should be noted that the metrics provided here are intended only to apply to management/performance and quality, not impact. Other methods can provide indicators to assist with understanding the impact of science, for instance, breadth of public knowledge and the socio-economic utility of that knowledge. The metrics described above are derived from the organization's publications, and therefore are bibliometric in nature. Non-publication aspects of an organization's management and quality are more appropriately assessed using more qualitative methods such as peer review. Evaluating impact, however, is beyond the capabilities of purely bibliometric methods because it involves measuring factors beyond the publications themselves.

References

2010. GPRA Modernization Act of 2010. 111th Congress, H.R. 2142. Retrieved from http://www.gpo.gov/fdsys/pkg/BILLS-111hr2142enr/pdf/BILLS-111hr2142enr.pdf.

Adler NJ, Harzing A-W. 2009. When Knowledge Wins: Transcending the Sense and Nonsense of Academic Rankings. Academy of Management Learning and Education 8(1):72-95.

Albarran P, Crespo JA, Ortuno I, Ruiz-Castillo J. 2011. The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics 88(2):385-397. doi:10.1007/s11192-011-0407-9

Albarrán P, Ruiz-Castillo J. 2011. References made and citations received by scientific articles. Journal of the American Society for Information Science and Technology 62(1):40-49. doi:10.1002/asi.21448

Belter CW. 2013. A bibliometric analysis of NOAA's Office of Ocean Exploration and Research. Scientometrics 95(2):629-644. doi:10.1007/s11192-012-0836-0

Borner K, Chen CM, Boyack KW. 2003. Visualizing knowledge domains. Annual Review of Information Science and Technology 37:179-255. doi:10.1002/aris.1440370106

Bornmann L. 2011a. Mimicry in science? Scientometrics 86(1):173-177. doi:10.1007/s11192-010-0222-8

Bornmann L. 2011b. Scientific Peer Review. Annual Review of Information Science and Technology 45:199-245. doi:10.1002/aris.2011.1440450112

Bornmann L, Daniel HD. 2008. What do citation counts measure? A review of studies on citing behavior. Journal of Documentation 64(1):45-80. doi:10.1108/00220410810844150

Bornmann L, de Moya Anegón F, Leydesdorff L. 2012. The new Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. Journal of Informetrics 6(2):333-335. doi:10.1016/j.joi.2011.11.006

Bornmann L, de Moya Anegón F, Mutz R. 2013. Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? Journal of the American Society for Information Science and Technology:n/a-n/a. doi:10.1002/asi.22923

Boyack KW, Borner K. 2003. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. Journal of the American Society for Information Science and Technology 54(5):447-461. doi:10.1002/asi.10230

Boyack KW, Jordan P. 2011. Metrics associated with NIH funding: a high-level view. Journal of the American Medical Informatics Association. doi:10.1136/amiajnl-2011-000213

Boyack KW, Klavans R. 2010. Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? Journal of the American Society for Information Science and Technology 61(12):2389-2404. doi:10.1002/asi.21419

Boyack KW, Klavans R, Borner K. 2005. Mapping the backbone of science. Scientometrics 64(3):351-374. doi:10.1007/s11192-005-0255-6

Butler L. 2003. Modifying publication practices in response to funding formulas. Research Evaluation 12(1):39-46. doi:10.3152/147154403781776780

Chen Y, Börner K, Fang S. 2013. Evolving collaboration networks in Scientometrics in 1978–2010: a micro–macro analysis. Scientometrics 95(3):1051-1070. doi:10.1007/s11192-012-0895-2

Costas R, van Leeuwen TN, van Raan AF. 2011. The "Mendel syndrome" in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists. Scientometrics

89(1):177-205. doi:10.1007/s11192-011-0436-4

Derrick GE, Pavone V. 2013. Democratising research evaluation: Achieving greater public engagement with bibliometrics-informed peer review. Science and Public Policy 40(5):563-575. doi:10.1093/scipol/sct007

Druss BG, Marcus SC. 2005. Tracking publication outcomes of National Institutes of Health grants. American Journal of Medicine 118(6):658-663. doi:10.1016/j.amjmed.2005.02.015

Eom YH, Fortunato S. 2011. Characterizing and Modeling Citation Dynamics. PLoS ONE 6(9):e24926. doi:10.1371/journal.pone.0024926

Garfield E. 1979. Citation Indexing. Its theory and application in science, technology and humanities. New York: Wiley.

Glanzel W. 2001. National characteristics in international scientific co-authorship relations. Scientometrics 51(1):69-115. doi:10.1023/A:1010512628145

Haeffner-Cavaillon N, Graillot-Gak C. 2009. The use of bibliometric indicators to help peer-review assessment. Archivum Immunologiae Et Therapiae Experimentalis 57(1):33-38. doi:10.1007/s00005-009-0004-2

Huang Z, Chen HC, Li X, Roco MC. 2006. Connecting NSF funding to patent innovation in nanotechnology (2001-2004). Journal of Nanoparticle Research 8(6):859-879. doi:10.1007/s11051-006-9147-9

Huang Z, Chen HC, Yan LJ, Roco MC. 2005. Longitudinal nanotechnology development (1991-2002): National Science Foundation funding and its impact on patents. Journal of Nanoparticle Research 7(4-5):343-376. doi:10.1007/s11051-005-5468-3

Jimenez-Contreras E, Lopez-Cozar ED, Ruiz-Perez R, Fernandez VM. 2002. Impact-factor rewards affect Spanish research. Nature 417(6892):898-898. doi:10.1038/417898b

Kessler MM. 1963. Bibliographic coupling between scientific papers. American Documentation 14(1):10-25. doi:10.1002/asi.5090140103

Kosecki S, Shoemaker R, Baer C. 2011. Scope, characteristics, and use of the U.S. Department of Agriculture's intramural research. Scientometrics 88(3):707-728. doi:10.1007/s11192-011-0359-0

Kostoff RN, Geisler E. 2007. The unintended consequences of metrics in technology evaluation. Journal of Informetrics 1(2):103-114. doi:10.1016/j.joi.2007.02.002

Lane J. 2010. Let's make science metrics more scientific. Nature 464(7288):488-489. doi:10.1038/464488a

Lane J, Bertuzzi S. 2011. Measuring the Results of Science Investments. Science 331(6018):678-680. doi:10.1126/science.1201865

Lee CJ, Sugimoto CR, Zhang G, Cronin B. 2013. Bias in peer review. Journal of the American Society for Information Science and Technology 64(1):2-17. doi:10.1002/asi.22784

Leydesdorff L. 2008. Caveats for the use of citation indicators in research and journal evaluations. Journal of the American Society for Information Science and Technology 59(2):278-287. doi:10.1002/asi.20743

Leydesdorff L, Bornmann L, Mutz R, Opthof T. 2011. Turning the tables on citation analysis one more time: Principles for comparing sets of documents. Journal of the American Society for Information Science and Technology 62(7):1370-1381. doi:10.1002/asi.21534

Liang L, Rousseau R, Zhong Z. 2013. Non-English journals and papers in physics and chemistry: bias in citations? Scientometrics 95(1):333-350. doi:10.1007/s11192-012-0828-0

Liebow E, Phelps J, Van Houten B, Rose S, Orians C, Cohen J, Monroe P, Drew CH. 2009. Toward the Assessment of Scientific and Public Health Impacts of the National Institute of Environmental Health Sciences Extramural Asthma Research Program Using Available Data. Environmental Health Perspectives 117(7):1147-1154. doi:10.1289/ehp.0800476

Lyubarova R, Itagaki BK, Itagaki MW. 2009. The Impact of National Institutes of Health Funding on US Cardiovascular Disease Research. PLoS ONE 4(7). doi:10.1371/journal.pone.0006425

Mahlck P, Persson O. 2000. Socio-bibliometric mapping of intra-departmental networks. Scientometrics 49(1):81-91. doi:10.1023/a:1005661208810

Martin BR. 1996. The use of multiple indicators in the assessment of basic research. Scientometrics 36(3):343-362.

Melin G, Persson O. 1996. Studying research collaboration using co-authorships. Scientometrics 36(3):363-377. doi:10.1007/bf02129600

Moed H. 2007. The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. Science and Public Policy 34(8):575-583. doi:10.3152/030234207x255179

Moed H. 2008. UK Research Assessment Exercises: Informed judgments on research quality or quantity? Scientometrics 74(1):153-161. doi:10.1007/s11192-008-0108-1

Moed H, De Bruin R, Van Leeuwen T. 1995. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. Scientometrics 33(3):381-422. doi:10.1007/bf02017338

National Research Council. 2012. Best Practices in Assessment of Research and Development Organizations. Panel for Review of Best Practices in Assessment of Research and Development Organizations; Laboratory Assessments Board; Division on Engineering and Physical Sciences. National Academies Press.

National Science Board. 2012. Science and Engineering Indicators 2012. Arlington VA: National Science Foundation (NSB 12-01).

Office of Management and Budget. (2009). OMB Memorandum M-10-01. Executive Office of the President. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-01.pdf.

Office of Management and Budget. (2010). OMB Memorandum M-10-24. Executive Office of the President Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-24.pdf.

Office of Management and Budget. (2013). OMB Circular No. A-11: Preparation, Submission, and Execution of the Budget. Executive Office of the President. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/a11_current_year/a11_2013.pdf.

Opthof T. 1997. Sense and nonsense about the impact factor. Cardiovascular Research 33(1):1-7.

Redner S. 1998. How popular is your paper? An empirical study of the citation distribution. European Physical Journal B 4(2):131-134. doi:10.1007/s100510050359

Redner S. 2005. Citation statistics from 110 years of Physical Review. Physics Today 58(6):49-54.

Rosas SR, Kagan JM, Schouten JT, Slack PA, Trochim WMK. 2011. Evaluating Research and Impact: A Bibliometric Analysis of Research by the NIH/NIAID HIV/AIDS Clinical Trials Networks. PLoS ONE 6(3):e17428. doi:10.1371/journal.pone.0017428

Schubert A, Braun T. 1990. International collaboration in the sciences 1981–1985. Scientometrics 19(1):3-10. doi:10.1007/bf02130461

Seglen PO. 1992. The skewness of science. Journal of the American Society for Information Science 43(9):628-638. doi:10.1002/(sici)1097-4571(199210)43:9<628::aid-asi5>3.0.co;2-0

Seglen PO. 1997. Why the impact factor of journals should not be used for evaluating research. British Medical Journal 314(7079):498-502.

The Royal Society. 2011. Knowledge, networks and nations. London: Elsevier.

van Dalen HP, Henkens K. 2012. Intended and Unintended Consequences of a Publish-or-Perish Culture:

A Worldwide Survey. Journal of the American Society for Information Science and Technology 63(7):1282-1293. doi:10.1002/asi.22636

van Leeuwen TN, Moed HF, Tijssen RJW, Visser MS, van Raan AFJ. 2001. Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics 51(1):335-346.

van Leeuwen TN, Visser MS, Moed HF, Nederhof TJ, van Raan AFJ. 2003. Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. Scientometrics 57(2):257-280.

van Raan AF, van Leeuwen TN, Visser MS. 2011. Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. Scientometrics 88(2):495-498. doi:10.1007/s11192-011-0382-1

van Raan AFJ. 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics 36(3):397-420.

Van Raan AFJ. 2005. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics 62(1):133-143. doi:10.1007/s11192-005-0008-6

Van Raan AFJ. 2006. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics 67(3):491-502. doi:10.1556/Scient.67.2006.3.10

Von Bertalanffy, L. 1950. An outline of general system theory. British Journal for the Philosophy of Science 1:134-165. doi:10.1093/bjps/I.2.134

Wagner CS, Leydesdorff L. 2005. Network structure, self-organization, and the growth of international collaboration in science. Research Policy 34(10):1608-1618. doi:10.1016/j.respol.2005.08.002

Wagner CS, Roessner JD, Bobb K, Klein JT, Boyack KW, Keyton J, Rafols I, Börner K. 2011. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. Journal of Informetrics 5(1):14-26. doi:10.1016/j.joi.2010.06.004

Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AF. 2011a. Towards a new crown indicator: an empirical analysis. Scientometrics 87(3):467-481. doi:10.1007/s11192-011-0354-5

Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AFJ. 2011b. Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics 5(1):37-47. doi:10.1016/j.joi.2010.08.001

Weingart P. 2005. Impact of bibliometrics upon the science system: Inadvertent consequences? Scientometrics 62(1):117-131. doi:10.1007/s11192-005-0007-7

Yang J, Vannier MW, Wang F, Deng Y, Ou F, Bennett J, Liu Y, Wang G. 2013. A bibliometric analysis of academic publication and NIH funding. Journal of Informetrics 7(2):318-324. doi:10.1016/j.joi.2012.11.006

Youtie J, Kay L, Melkers J. 2013. Bibliographic coupling and network analysis to assess knowledge coalescence in a research center environment. Research Evaluation 22(3):145-156. doi:10.1093/reseval/rvt002

Zoss AM, Borner K. 2012. Mapping interactions within the evolving science of science and innovation policy community. Scientometrics 91(2):631-644. doi:10.1007/s11192-011-0574-8