Automating Scientometric Analyses for Large Research Institutions

Austin Richards[1], Sean J. Paul[1]
1. Office of Research and Development, U.S. Environmental Protection Agency

Austin Richards Richards.Austin@epa.gov | 202-564-3263

Abstract

The principal tools of Scientometrics, the academic discipline concerned with evaluating scientific literature performance, often incorporate citation counts. These tools include journal impact factor, the H-index, and emerging 'alt-metric' scores that quantify social media and non-academic mentions and references. It is commonplace to use these indicators to gauge the performance of individual scientists and publications, but scientometric indicators also have the potential to serve as evaluative tools for large research institutions. We gathered and analyzed 6 years of scientometric data associated with papers published by the US EPA's Office of Research and Development (ORD), and developed a tool for low effort automated continual data gathering. Based on the enterprise-level data collected, we developed performance metrics extending beyond simple citation counts.

Common Scientometric Indicators

Using scientometric indicators to quantify the scientific impact of a body of research is commonplace in academia and the public and private sectors. Citation counts have been shown to strongly correlate with other measures of research notoriety.[a] Further, across the US Federal Government, the performance of a federal scientist is determined in part through an evaluation of their scientific contributions, and citations are specifically considered as an indicator of competence and productivity.[b]

Indicator	Common Unit of Analysis	Formula Description
H-Index	Individual Scientist	N number of publications with at least N citations received
Journal Impact Factor (JIF)	Scientific Journal	Yearly average number of citations that articles published within the past 2 years (in a given journal) received
Alt-Metric Score	Individual Publication	Weighted index value incorporating counts of publication mentions in news, social media, Wikipedia, and other popular sources

Table 1. Common Scientometric Indicators for Research Impact Evaluation

Methods

ORD researchers published over 4,500 peer-reviewed articles between 2013 and 2019 and publish ~50 more articles each month. ORD's scientists are active in a broad range of disciplines from ecology to social sciences. Using an Agency subscription to Web of Science (WoS), and open-access software packages in R and Python, we wrote a web script which queries the Web of Science database and retrieves an organized profile of all scientific literature which has cited ORD's work. Our data input for this system is the Digital Object Identifiers (DOIs) of each of ORD's publications, which are gathered internally. Because our Web of Science query is executed programmatically, the citation profile we built can be refreshed and updated on a regular time interval. A separate supplemental web script we developed also retrieves alt-metric indicators associated with ORD's published work. Finally, using WoS journal subject characterizations we constructed a detailed profile of ORD's research output, and are able to conduct within subject analytical comparisons.

Results

Figure 1.

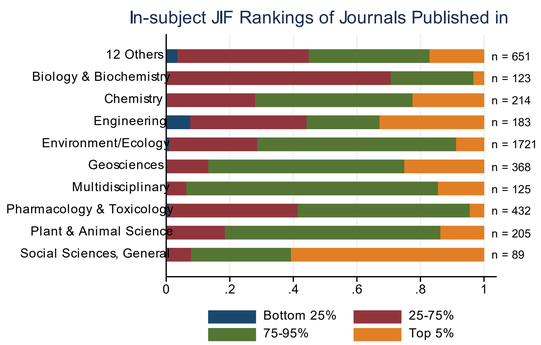
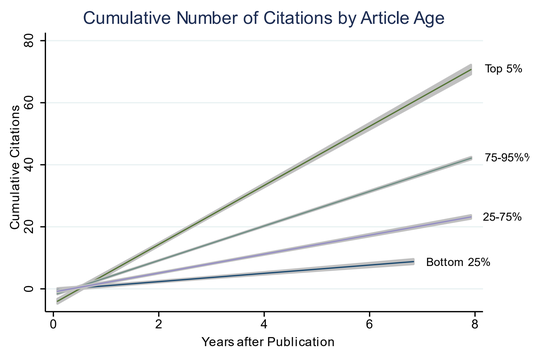


Figure 2.



New Performance Metrics

To account for the heterogeneity in topic area and age of ORD's published works, our data collection system supports continuous measurement of performance metrics that incorporate age weighting and within subject comparisons, such as:

- Average, in-subject journal impact factor ranking of journals ORD has published in.
- Age weighted citation rates of ORD's publications vs in-subject norms.
- Average alt-metric scores of ORD publications vs in-subject norms.

Findings

Between 2013 and 2019, ORD most frequently published in high-profile journals (top 25% in-subject JIF) in Geoscience, Multidisciplinary, and Social Sciences works (Figure 1). During the same time period, ORD's articles that were published in higher ranked journals accumulated higher numbers of citations in the years proceeding publication (Figure 2). This provides some evidence that in-subject JIF ranking provides a reliable early indicator of article impact in its field.

Tools

In addition to an Agency subscription to Web of Science as part of Clarivate Analytics which provided an API key, this evaluation utilized the WoS Python package and the following R packages: rAltMetric, tidyverse, reticulate.

Discussion and Limitations

The tool we developed enables low effort continuous tracking of key scientometric indicators of ORD's published research. Leveraging information on journal subject profiles from Web of Science, the performance metrics we developed and are able to continuously track extend beyond simple citation counts. Our work provides an example of how scientometrics can be leveraged to evaluate a large volume of heterogeneous research, such as the full body of work produced by a large institution.

Scientometrics rely heavily on citation counts, which offer an imperfect proxy of research impact. Some citations exist as simple perfunctory nods to related research, or in the worst case, a call out of invalid methodologies.[c] Still, these indicators do provide some informational utility.[d] All citation counters rely upon imperfect software, so inevitably some citations are either missed or double counted. Due to rate limiting of the WoS API, code processing for gathering citation information of thousands of articles at once is time intensive.

Sources:
a: Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16:317–323.
b: https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/functional-guides/gsresch.pdf
c: Cronin, B. (1982), "Norms and functions in citation – the view of journal editors and referees in psychology", Social Science Information Studies, Vol. 2, pp. 65-78.
d: Bornmann L, Daniel H-D (2008) What do citation counts measure: A review of studies on citing behavior. J Doc 64(1):45–80.