A bibliometric approach for managing institutional and personal scientific investments, with examples from the biological sciences.

Kevin Winker¹

¹ University of Alaska Museum, 907 Yukon Drive, Fairbanks, Alaska USA, 99775.

kevin.winker@alaska.edu

Abstract

Self assessment should include asking ourselves how we might allocate our investments to enhance the impact of our publication portfolio. I develop an easily implemented scientific investment portfolio analysis tool based on citations. Using Google Scholar data, I provide examples at individual and institutional levels for two cases in the biological sciences. Visualizing these data in three dimensions reveals striking degrees of structure and variation in how investments have been made and in how they have performed among subdisciplines or scientific market sectors. Legacy and time-corrected performance also provide contrasting views. This approach provides a quantitative way to assess market sectors in relation to each other in a way that should be broadly useful in planning future scientific investments for individuals, departments, or institutions.

Keywords Biological collections, citations, Google Scholar, planning, self-assessment, scientific market sectors.

Introduction

Using citations as a measure of scientific impact has become a popular and widespread method to assess aspects of performance at individual, departmental, and institutional levels. Although citations are an imperfect and incomplete measure of overall performance, they are nevertheless readily quantified and useful to examine one dimension of impact: how the publications of a person or unit are being used by fellow scientists. The utility of this metric in assessment is heavily debated, particularly because it does not accurately reflect the full contributions that scientists make, but these debates usually focus on external evaluations (e.g., San Francisco Declaration on Research Assessment 2013, Eyre-Walker and Stoletzki 2013). Less frequently discussed is how they might be used for internal assessment. As individuals or institutions we should routinely ask ourselves how we could spend time and money effectively in developing and producing scientific output. One way to pose such a question might be: How should we allocate our investments to enhance the impact of our publication portfolio? Here I develop a scientific investment portfolio analysis tool based on citations that is easy to implement. I also provide examples at individual and institutional levels for two cases in the biological sciences.

Materials and methods

This method is based on citation data from Google Scholar profiles. Unfortunately, Google Scholar does not have an application program interface (API), and it prohibits programs from scraping these data, making it a difficult target for automated data harvest. Here I used two Google

Scholar profiles, my own and one that we created for the University of Alaska Museum bird collection (UAM Birds; Winker and Withrow 2013). I manually harvested the data as follows. Sign into your Google Scholar profile. When looking at the list of citations, click on the Title box to check them all, then on Export, CSV, Export all my articles, and Export to obtain a spreadsheet file containing each publication and the fields Authors, Year, Title, Publication, etc. Next, manually retrieve the total citation numbers for these and enter them into the spreadsheet (the order of output does not match that of citation frequency). An example of the structure of these results is given in Supplemental Table S1.

Next, the individual publications are assessed and allocated into disciplines or subdisciplines that you create that make internal sense—i.e., areas recognizable as ‘market sectors’ to the individual or institution doing the self-assessment. The goal is to obtain the number of citations for each research area and the average date of publication for articles in that area. In my examples I chose not to double count, so interdisciplinary works had to go into a single category or into ‘miscellaneous.’ Each of these subdisciplines or areas becomes a column in the spreadsheet to the right of the main dataset, with an additional column for date of publication, and these are populated from the main dataset as each row (publication) is categorized (Supplemental Table S1). At the bottom, total citations and average date of publication for each research area are tallied (Supplemental Table S1).

The final data manipulation steps involve taking these summary values and adding a correction for time. While it is useful to examine legacy performance (and this is included in these examples), it is also useful to consider that publications gain citations through time, and the rate at which sectors do this is also a good indicator of performance. A sector-based, time-corrected citation rate (Supplemental Table S1, sheet Graphing) is calculated for each sector as: $(\text{citations/publications})/(\text{current year minus average date of the sector's publications})$. These values are then graphed in three dimensions to visualize the results. The dimensions are Y) number of publications, X) average date of publication, and either Z_a) total citations or Z_b) time-corrected citation rate, with the third dimension given as size of sector (Supplemental Table S1, sheet Graphing).

Results

Visualizing these data in three dimensions provides insight into the legacy and time-corrected performance among investments in the various scientific ‘market sectors’ relative to each other (Figs. 1, 2). In the upper panels (Figs. 1A and 2A), the bubble sizes show the cumulative citations per market sector and are the raw data on how past investments have built total citations and the overall *h*-index. In the lower panels (Figs. 1B and 2B), the bubble sizes show time-corrected citation rates (average citations per publication per year relative to sector’s average publication age) by market sector and help us understand relative sector performance through time (Figs. 1, 2). An aggressive investment strategy would focus more resources on the larger bubbles in the lower panels, recognizing that the per-publication citation rates there are likely to continue to be higher, at least over the shorter term during which these sectors have been measured in these datasets.

Temporally, an increasingly interdisciplinary portfolio is apparent in Fig. 1, with large recent investments (numbers of papers) in genetics and phylogeography. This reflects the comparatively recent development of a genomic resources component to the UAM bird collection. Researchers capitalizing on that institutional investment have outstripped the more historic uses of the collection in faunistics and biodiversity, and the relative rates of citations in the two sectors suggest that further

institutional investments in developing this aspect of the collection will probably continue to have a strong impact (Fig. 1).

Temporal shifts are also evident in Fig. 2, with a move away from avian ecology and into biodiversity, genetics, and disease research. These were conscious decisions on my part over time (PhD 1991), but this graphic portrayal provides useful insights. Although I continue to invest in an interdisciplinary portfolio, since developing this tool I have put more investment into disease research.

Figure 1. Graphic portrayal of citation data by scientific 'market sector' in terms of numbers of papers, average date of publication, and A) total citations (relative size of spheres) and B) sector-based, time-corrected citations (relative size of spheres). The Google Scholar profile portrayed is UAM Birds.

Figure 2. Graphic portrayal of citation data by scientific ‘market sector’ in terms of numbers of papers, average date of publication, and A) total citations (relative size of spheres) and B) sector-based, time-corrected citations (relative size of spheres). The Google Scholar profile portrayed is that of Kevin Winker.

Conclusions

This data visualization approach reveals striking degrees of structure and variation in how investments have been made and in how they have performed among subdisciplines or scientific market sectors (Figs. 1 & 2). This perspective should prove invaluable for planning future investments. Of course, as they say about stock markets, past performance does not predict future returns. And we know that in biology scientific sectors can have inherently different citation magnitudes (e.g., Mingers and Meyer 2017). For example, taxonomic publications on average garner

fewer citations than those in medicine.

One important way in which this management tool differs from monetary investment approaches, which I have used only as an analogy, is that we cannot retroactively move investments about. Once time and money have been put into a project, they are relatively fixed. Other factors, such as journal impact factors and collaborator networks, can also be important considerations (Winker 2011), as can funding opportunities. Exceptions to the relatively fixed nature of historic investments do exist, and with forward planning can enhance a future portfolio. For example, for specimen-based biologists there are opportunities to use specimens archived from historic projects in new and different ways. And of course historic datasets might also be applied in new areas.

Time-corrected citation rates can show striking changes in relative sizes among sectors when compared to legacy, raw-citation values (e.g., contrasting upper and lower panels in Figs. 1 & 2). One caution in interpreting large proportional jumps in the time-corrected values is that attention to publications shows early spikes and then wanes over time (Parolo et al. 2015). Thus, recent activity can disproportionately affect apparent importance, particularly in sectors that don't have a long legacy of publications to ameliorate these effects. However, recency alone is not likely responsible for the relative importance of time-corrected rates in the sector Disease/Contaminants (Figs. 1B, 2B) given more modest shifts in Genetics/Phylogeography. These differences probably reflect genuine sector-specific differences in audience sizes. Numerous additional, underlying factors might also be considered, such as journal half lives, coauthors, scientific novelty (per Stephan et al. 2017), etc.

Although we are usually aware of which papers and areas of research in our portfolios have been important from the perspective of numbers of citations, this tool provides a quantitative way to assess market sectors in relation to each other in a way that should be broadly useful in planning future scientific investments for individuals, departments, or institutions.

Acknowledgments

I thank Thomas Buhrmann for his program (<https://buhrmann.github.io/scholar.html>) and help in attempts to obtain Google Scholar data automatically (although we were ultimately unsuccessful). This research did not receive any grant funding.

References

- Eyre-Walker, A., & Stoletzki, N. 2013. The assessment of science: the relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biology* 11:e1001675. <https://doi.org/10.1371/journal.pbio.1001675>
- Mingers, J., & Meyer, M. 2017. Normalizing Google Scholar data for use in research evaluation. *Scientometrics* 112:111-1121.
- Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. 2015. Attention decay in science. *Journal of Informetrics* 9:734-745.

San Francisco Declaration on Research Assessment (DORA). 2013. <http://am.ascb.org/dora/>

Stephan, P., Veuglers, R., & Wang, J. 2017. Blinkered by bibliometrics. *Nature* 544:411-412.

Winker, K. 2011. In scientific publishing at the article level, effort matters more than journal impact factors: hard work and coauthors overshadow journal venue in acquiring citations. *BioEssays* 33:400-402.

Winker, K., & Withrow, J. J. 2013. Small collections make a big impact. *Nature* 493:480.