General principles for the responsible use of bibliometric indicators

August 2020
bibliometrics@ucl.ac.uk

The UCL bibliometrics policy sets out some principles for the use of citation metrics in research assessment at UCL. As part of the overall guidance, this paper sets out some general principles for the responsible use and interpretation of citation metrics.

Whatever your purpose for using metrics, it is always important to consider the inherent limitations in citation indicators - it is easy to produce misleading figures without intending to.

Here, we highlight some common factors known to create biases in bibliometrics. You may encounter these during your analysis or interpretation, and where appropriate you should take steps to account for them.

This is not an exhaustive list, and in some cases your purposes and analyses will be unique to you, and so it is your responsibility to ensure you have assessed the limitations of the metrics you intend to use and consider whether other forms of assessment may be better suited.

- **Disciplines**: The traditional sources of bibliometric data have focused on indexing journal articles. They may therefore be less useful in disciplines that have a greater diversity of research output types, such as the arts, humanities, social sciences – where books or edited collections are more common - or computing science and engineering, which have a greater focus on conference papers. Some useful questions to ask yourself are: "Am I being comprehensive in collecting research outputs?", "Have I omitted any particular research output type based on my data source?", "Do journal articles reflect research output for my field/purposes?".

  Further, the disciplines that do concentrate on journal articles may still have significantly different publication and citation practices. Some indicators can be normalised to allow comparisons across disciplines, but this is only ever approximate, and may be challenging for multidisciplinary work.

  Consider the scale of any normalized metrics as applied to disciplines. Most normalized metrics provided by bibliometrics data sources are calculated at the journal level or higher. Higher level discipline groupings are broad, and have more limited use at individual research or outputs levels. And normalisations by discipline do not usually account for variation in approach. E.g., even within a narrow subfield, an experimental researcher may have very different publishing and citation practices than a theoretical researcher.

  Different disciplines also have different approaches to collaboration and multiple authorship. But because authorship is the primary means to attribute credit and contribution to journal articles, it may be appropriate to consider discipline specific relationships between authorship and research contribution when assessing research or researchers among disciplines.

  Be alert to the mix of publications in your data; be careful when drawing comparisons between different disciplines; be aware that useful data will not always be available for all disciplines; and use normalised indicators where possible.

- **Publication type:** Review articles will generally get more citations than original research. It is crucial to consider the effect this may have on citation-oriented bibliometric analyses; and omit reviews where appropriate.

  When calculating citation figures for a researcher, try to filter out non-research material such as editorials and book reviews, as these will skew the figures. For some research outputs, such as practice-based material or policy documents, it is unlikely that any citation analyses would be productive and an alternative method of assessment is encouraged.

- **Different meanings of citation:** Citations do not always signify the same thing. A paper may be cited as the fundamental basis of a later work, or as a cursory reference in a long list of

loosely related prior research; either of these will show up as a single citation in a database and be counted in the same way with the same perceived value. Negative citations also exist – a particularly contentious paper may be frequently cited in order to dismiss its findings, or merely to remark upon the controversy.

It is not usually possible to avoid counting negative citations, and in some fields they may be more frequent than others. Be aware of their existence, and look critically at any strange outliers.

- **Multiple authors:** Where a document has multiple authors, most databases and indexes will attribute a publication and associated citation data to all authors. This means a single publication will count once for each author regardless of whether they were the sole author or somewhere in a long list of authors; and once for each author's institution regardless of whether the institution contributed one or more authors. This can give misleading results, particularly in highly collaborative fields.

  Fractional counting is possible, but relies on assumptions about contribution, such as attributing equal portion of an output to all authors, or weighting based on author order. Yet these conventions can vary substantially among individual papers, journals, and fields.

  Currently, there is no standard or accepted way to address multi-authored research outputs for bibliometric analyses. Without a fixed standard, shared authorship outputs must be interpreted cautiously in bibliometric analyses, whether attempts have been made to account for contribution or not.

  Where relevant and possible, consider authorship practices across the fields of interest, acknowledging variation in practice and subjective decisions within the field. If you are only looking at a small number of papers, it may be practical to look at author contribution statements if they are present.

- **Time since publication**: Citations continually accrue over time, so older papers will tend to have more citations than newer ones. Similarly, older researchers usually have more papers and have had more time to accrue citations to them, so an indicator like the h-index will be misleading for researchers at different stages of their career.

  Where possible, compare papers against those of a comparable age, use a fixed time period after publication, or estimate of citations per year.

- **Basic principles of good data and statistical analysis and interpretation:** When using and interpreting metrics, particularly those that result from calculations or aggregations, ensure you have sufficient knowledge of fundamental data analysis practice and basic statistical principles to undertake the desired work; and do not ask those that are not skilled in these areas to make decisions based on such metrics and statistics without clear interpretation guidance. Common issues include the following:

  Do not apply aggregate level metrics to individual subjects, and vice versa; such as incorrectly assess the quality of individual papers based on the impact factor of the publishing journal.

  Understand composite metrics expressed as single values (i.e. single value metrics based on calculation among other metrics), fail to capture the complexities, nuances, and multidimensional attributes of the underlying data. And still, they often miss or cannot encompass or account for all of the mediating factors for the characteristic being estimated, particularly at the individual level. This is why UCL recognizes the h-index is not valid - and metrics in general should be avoided - when assessing individual researchers, as they do not account for diversity in research (sub)disciplines and individually variable attributes (e.g., niche research questions, career stage, personal circumstances, and teaching/administrative commitments; see below for more on Author Background).

  Aggregated metrics, such as calculated means, are sensitive to sample sizes and give no information on variability. Hence, these can be unreliable if based on too few observations, as it's hard to distinguish whether the mean or aggregated metric represents a genuine indication or trend, or is the result of chance or subject to outliers with high leverage. At a

minimum, the number of observations used to calculate an aggregate metric should accompany written or graphical representations of the metrics, as well as at least one estimate of dispersion, variation or distribution (e.g., standard deviation, error bars) for means.

- **Author background**: The personal background of the author can affect both their publication rates and their citation rates. Full-time researchers will tend to produce more papers than those on part-time contracts, those who have recently had career breaks (for example, through illness, or family or care commitments), or those with substantial non-research responsibilities (teaching, administration, etc).

  Individual factors (such as gender, ethnicity, age, and language) may also lead to existing large-scale systemic biases being reflected in (or even reinforced by) the citation indicators. For example, gender biases, where male researchers are cited more often than female (known as the Matilda Effect), are well documented. Further, a known “Matthew effect” – the rich get richer – demonstrates the more citations tend to be accrued by more prominent, well-established, researchers.

  These are hard to normalise for, and as such indicators based on individuals should be treated with caution.

- **Data sources**: Different citation databases will give different results, because they index a different range of material in different ways. Some have better coverage of monographs, reports, and conference proceedings than others; some omit specific journals. Some have better coverage of non-English language sources than others.

  Comparisons should always be based on the same data source, and where possible using data gathered at the same time. Always be cautious of benchmarking against citation indicators from an unclear source.

- **New and alternative metrics**: There has been a sharp growth in recent years in various commercial "altmetrics" services. These often use similar source data (eg number of tweets or download figures) but interpreted and presented in different ways. Depending on what indicators are used, they can show scholarly interest (eg Mendeley bookmarking), media interest (eg news stories), or public interest (eg social media activity). They can also be used to identify the use of research in policy documents or other official publications which may not appear in the conventional citation databases. Spikes in activity may come if a piece of work is particularly contentious, timely, or simply on a topic that catches the public imagination. It is harder to gather standardised and comprehensive new metrics than it is traditional citation data.

  In general, it is best to treat figures from these metrics as broad indicators – high activity tells us that there is something interesting there, but the details should be examined before drawing conclusions. They should never be used to quote a single numeric “score” for ranking a paper or author.