Contents lists available at ScienceDirect

Journal of Informetrics

Journal of INFORMETRICS
An International Journal

Paving the way or pushing at open doors? A comment on Abramo and D'Angelo "Farewell to size-independent indicators"

CrossMark

1. Introduction

The article by Abramo and D'Angelo (2016; hereafter AA) is an interesting piece in the debate about indicators, but it is a bit of a caricature. Firstly, a few words on the context.

If we define scientometrics as the study of science actors and networks, and "output bibliometrics" as the particular networks involving knowledge media, it is the nature of large disciplines with powerful conceptual frames, such as mainstream economics and evolutionary economics (as it would be for the various strong streams of sociology), to take over bibliometrics as soon as they address technology and science. For economics, science output came later in the game than patents, since scientific knowledge was basically considered as a public good (among others Arrow, 1962) albeit gradually appearing as less pure. Stephan for example, considers it as a mix of public goods in the information process and rival goods, prestige and priority, in the scientific competition (Stephan, 2012). Once publication was recognized as a legitimate object, economics thinking stimulated reflections not only on the cost side of research, but also on benefits and incentives of academic publication (on the value of publications and citations, a landmark is Diamond, 1986). More generally, economics of science (as sociology of science) rightly claims science networks, the core of scientometrics, as important objects, with two implications:

- Output bibliometricians should remain modest and prudent, and welcome strong concepts and the toolbox coming from economics. On the policy side, many bibliometricians have devoted much time to warn against the limitations of bibliometric measures, to emphasize their properties and their conditions of use, to stress the dangers of misusage, and to insist on the traps of one-figure descriptions.
- Conversely, authors with another background should be cautious in the criticism of bibliometric measures considered independently from their historical context or the question of sources and data.

We shall focus here on archetypal indicators coined by evaluative bibliometrics, the size-independent measures like impact which attract AA's wrath. In fact, we shall not single out the mean normalized citation score (MNCS) but rather the archetypal form of impact which is the general target. For the sake of discussion, we will admit that "production" refers to a count of output (citations, publications, patents, etc.) and that "productivity" refers to a count of output by unit of input such as human resources – considered here – or economic resources. At the individual level, production and productivity are equivalent with unit staff input, but in practice alignment of output and input (e.g. time allocation) may be tricky, the question of multiple types of input and input at the group level has inspired global approaches of university productivity using tools such as data envelopment analysis, a complement to conventional production functions. We shall not enter in technical points such as data standardization, relative indicators and normalization, which are studied at length in the literature.

2. Strictly bibliometrics: a tale of two figures

Evaluative bibliometrics set up in the seventies, in its mainstream, relied on a few pillars:

- Focus on meso or macro-level indicators, with a worldwide view of science and on policy applications, rather than on the individual scientist's level. This is in compliance with de Solla Price's warning to "*not fix one's gaze on a specific molecule called George*" (de Solla Price, 1963). The calculation in terms of impact is primarily actor-scale independent, so that in principle one can calculate data for groups (institutions, regions, journals, etc.) with individuals as a limit case. An advantage is that analysis at the aggregate level does not require a detailed knowledge of individual assignments, as far as the targeted

level of analysis is correctly identified in the database affiliation.[1] However, the border of individual evaluation was soon crossed, with the use of the journal impact factor in savage evaluation of individuals, leading to the booming of the h-index (Hirsch, 2005) responding to growing evaluation needs in institutions.

- Telling a tale of two measures: the core of output bibliometrics short-cuts external data (staff, funding) handled by general scientometrics and economics of science. Up to the present period, there has been a strong justification to this form of "arte povera": output data – publications and citations – pose a variety of methodological problems, but once the basis of evaluative bibliometrics has been laid, they paradoxically appeared more reliable than input data,[2] staff or funding, generated by accounting systems. Despite attempts by the OECD to standardize research data at the international level as soon as the early sixties with the first Frascati manual, it remained extremely difficult to compare staff data across nations, even within the EU for example; the same was true for research funding, with lasting controversies about the productivity of national research systems. Deprived from accurate and comparable input data, the modest approach of bibliometrics made a virtue out of a necessity. As mentioned below, there is now progress towards better identification of staff resources over the world scientific community.

- The two basic ingredients of output bibliometrics need some contextualization (selection and conditioning of data) and statistical elaboration, eventually resulting in a fireworks of indicators with various time-frame treatments, mathematical solutions for averaging and normalization, and pure or composite citation-publication measures. Let us forget this and focus only on archetypes. When we talk about macro or meso-level data (groups such as institutions, regions, nations, etc.), the historical target of evaluative bibliometrics, the volume of publications and citations obviously depend in some way on the scale of the analysis and the size of the entities. In contrast, the ratio of citations and publications, which is the archetype of a "bibliometric impact" measure, appears as primarily scale-independent or size-independent. In more practical terms, total publications and especially total citations could be seen as groups' "power" indicators, while impact could be interpreted as "performance". A current practice amongst indicator producers was then to publish impact indicators plus one or both of the associated power indicators: total volume of publications and/or citations.

The impact as described above is a baseline, which further allows the highlighting of scaling effects, especially increasing or decreasing returns of citations to the intermediary product "publications". The concept of impact remains valid for individuals and groups, although raising many technical questions: transformations of variables, different types of central values, and issues related to tails of citation distributions.

The tails issues in skewed citation distributions are addressed in different ways in those families of measures. The high tail may jeopardize indicators with excessive sensitivity to outliers, and the low tail suffers from the arbitrary delineation of most databases. Moreover the content of the tail is ambiguous. For example authors of great prestige may also like to publish popular science articles, which is detrimental to their classical average measure of impact. The truncation of the low tail has been recommended on original data beforehand (CWTS, OST-Paris, etc.) on various criteria. The irregularities in the high tail are addressed by a variant of impact measures, the profiles and scores measures. A basic one is the top percentile(s) count, where the size-independent element (impact) is replaced by a ratio like HCpub divided by publications, where HCpub denotes the count of publications within the highest percentile(s), or alternatively beyond some citation threshold. The corresponding "size dependent" indicator is HCpub, privileged by AA. This can be studied through several levels of visibility, giving profiles which can be summed up in global scores, with good properties albeit some arbitrariness in the weighting of citation classes. The composite h-index originally coined for the individual level brought an original solution to the problems of tails by implicitly ruling out both tails defined for each individual. Transformed measures (for example logarithmic transformation) may also be used, with a different rationale of averaging.[3]

The counterpart of relying on basic measures is to remain modest in the claims. One point on which AA hit the target is their point against some imprudent claims about "crown" indicators. The same might be said of making idols of h-index variants covering the entire alphabet. Quite a few of us, as bibliometricians, repeatedly voiced their concern about the "one-figure" culture in evaluation, and this is now part of several manifestos from science institutions and societies.

3. A tale of three measures

The central argument of AA is summarized in the sentence: "*Given two universities of the same size, resources and research fields, which one performs better: the one with 100 articles each earning 10 citations, or the one with 200 articles, of which 100 with 10 citations and the other 100 with five citations? A university with 10HCAs out of 100 publications, or the one with 15HCAs out of 200 publications?*" An important point is made here: '*given two universities of the same size, resources and research fields*'. The

---

[1] For example, if the institution's name is correctly captured by the database. In practice, the unification of such data is sometimes very poor, needing a time-consuming detour by individual assignment tasks when evaluation is at stake.

[2] At least in the mainstream of hard disciplines. This can be challenged in domains where data coverage and quality are poor, for instance a few disciplines in social sciences and humanities, and some parts of engineering and computer science.

[3] Significantly, the first historical impact measure, Garfield's journal impact factor, was criticized for its summarizing of skewed citation distributions, mainly because this indicator was misused for the evaluation of individual researchers. Actually, if the level of evaluation is the journal, the capability of rejecting low-expectation articles can be seen as a sign of efficiency of the journal's strategy. Quite different is the context of the evaluation of individuals or labs, where visibility on popular media in addition to academic journals may be part of the general policy, albeit at the expense of lower impact means.

problem is that this clause assumes prior information on input (size). In fact we should distinguish between study domains where information on staff is available and controllable, and study domains where it is not, and this is a frequent case in vast international comparisons. With reliable staff data, HCpub is a nice measure with respect to citation limits and classification issues, and we can easily derive size-independent indicators such as average individual count of HCpub. Without reliable staff data, the expectation of the mere number of HCpub being obviously sensitive to size, what interpretation can we make? Mean values (with a proper caution on tails) or proportion of HCpub may be "less bad" performance indicators than the mere count of HCpub, which however remains a good "power" indicator for groups.

Let us now consider productivity to staff, both for citations and publications. From an economics viewpoint, one could think of an analogy of citation with "values" and of publication with "volumes" and in this case impact would have the dimension of an average price. As published science is a public good, this analogy works poorly. Informally, citations are more acknowledgements of usage of some kind, depending on the sociological interpretation – real Mertonian use, economic utility or symbolic use – discussed at length in the literature.

Talking about groups, if we state that volume of publications and citations is expected to grow with the number of staff or the amount of funding all other things equal, it does not mean that the volume of publications is a valid proxy of more intuitive measures of size such as staff. A basic relation (here on means) is that staff citation productivity can be trivially written as the product of publication productivity and bibliometric impact

$$cit/R = (cit/pub) \times (pub/R) = impact \times (pub/R),$$

where cit/pub impact is a strictly bibliometric measure and pub/R is the staff publication productivity. Publications appear either as an intermediary resource in the production of citations, or as an intermediary product of staff. This ambiguity is reflected in terminology: strictly speaking, impact is by definition corrected for the size of the literature and primarily size-independent from this point of view, but of course "size" is usually meant as number of staff and not of papers. In this case, impact is logically size-indifferent, albeit of course empirically subject to scaling effects such as increasing or decreasing returns on group size.

Therefore impact appears as a component and not a substitute of the citation productivity of staff, and no bibliometrician would claim it is. This decomposition of citation productivity would be artificial if it did not highlight two moments of the scientific production: the production process of published research and the circulation of findings. The other way around, impact may be seen as the ratio of the two productivity indexes (citation and publication) based on an identical resource. The relation above also expresses that publication is equivalent to a size measure only in the case where publication productivity is considered constant in the area under scrutiny. In this case impact and citation productivity are proportional.

This primarily staff size-independent measure cit/R should be complemented by indicators such as total publications and/or citations, which reflect the power of a group. The latter can be trivially rewritten as cit = impact × pub to highlight these two moments.

The difference between bibliometric impact and citation productivity is evident, and needs to be mentioned only in special contexts. For example, talking about citation normalization, Zitt (2013) stated "*we have not discussed the related question of normalization of publication (or citation) productivity by domain. In other words, in the ratio citation/resources, decomposable in two factors, citation/publication and publication/resources, the normalization target considered here only addresses the first factor. The productivity ratios to staff or funding raise severe methodological and political issues, as well as problems of data availability*".

Now the above mentioned question remains: do we have reliable staff figures for a particular study? There are various steps by which data become reasonably comparable, institution-wide, region-wide, nation-wide, geopolitical region-wide and worldwide. A crucial point is that a reliable input measure should exist both for actors under scrutiny and the benchmark. This may be easier for focused surveys, say within a nation-discipline framework (e.g. Italian physicists, without external benchmark). If we wish to compare all Italian academics or all European physicists, the input data issue is already more serious.

By default, bibliometric studies often rely on a world baseline as the reference, embodied for example in "relative impact indicators" (the simpler form is the ratio to world average), since this is feasible for article-based indicators in the Web of Science and Scopus databases. This worldwide benchmark may become terribly difficult to collect for author-based productivity indicators, asking for a global unification/disambiguation of authors names, preferably on a permanent base, and more generally for a correct identification of publishing and non-publishing staff with comparable status over systems. This remains a big challenge of "data-demining" in operational bibliometrics. Online CVs, self-validated publication lists within databases (REPEC, Web of Science/Scopus or Google Scholar type), and the ORCID initiative, among others, promise good perspectives of measurement improvement. Other collaborative projects, especially the inter-university Aquameth project (Daraio et al., 2011) and projects at the national level (Norway, Brazil, etc.), have made steps in the same direction. As soon as data become reliable and comparable within and across nations and disciplines, there is no doubt that productivity data become an important tool in assessment and that performance comparisons based on total HCpub will become less dangerous.

4. Conclusion

To summarize, bidding farewell to impact measures without a fair trial is rather imprudent:

- Productivity analyses make sense at a given scale if there is strong evidence that input data are comparable, which is sometimes achieved in studies of universities, especially within a national system. Otherwise, the cure is worse than the disease and size-dependent measures are useful, of course, as power indicators but say very little about performance. The quest of unique researcher IDs is promising, but still ongoing.
- Even if data are comparable, discarding the concept of impact might be unwise, since publication productivity and impact address two complementary components and two distinct moments – albeit intertwined – of the research process.
- It remains that many bibliometric measures deserve critiques. None is perfect, none can be used alone. Furthermore, a compromise should often be sought between axiomatic soundness, social significance, and availability of data. The fact that some size-independent measures are not fully satisfactory cannot be a motive for condemning the very principle of size-independent characterization. The recurrent scaling issues and especially the question "is small beautiful in research?", within appropriate production functions schemes, ask for more elaborate arguments.

Eventually, the tale of three measures is a much richer story than the previous one, which already proved able to generate a cornucopia of indicators. However, both remain tales if they fail to capture the creativity of scientists at given moments of their personal trajectories and in a variety of disciplinary contexts. The dream of driving academic research by recipes borrowed from management techniques might turn into a nightmare.

Farewell to size-independent indicators, and come back soon?

References

Abramo, G., & D'Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics, 10*(2), 646–651.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In R. Nelson (Ed.), *The rate and direction of inventive activity: economic and social factors* (pp. 609–626). Princeton University Press. NBER.

Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., et al. (2011). The European university landscape: a micro characterization based on evidence from the Aquameth project. *Research Policy*, *40*(1), 148–164.

de Solla Price, D. J. (1963). *Little science, big science (foreword)*. Columbia University Press (chapter 1).

Diamond, A. M. (1986). What is a citation worth? *Journal of Human Resources*, *21*(2), 200–215.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Stephan, P. (2012). *How economics shapes science*. Harvard University Press (chapter 1).

Zitt, M. (2013). Variability of citation behavior between scientific fields and the normalization problem: the 'citing-side' normalization in context. *Collnet Journal of Scientometrics and Information Management*, *7*(1), 55–67.

Michel Zitt
Lereco (SAE2) U1134, INRA, Nantes, France (rtd)
E-mail address: mzitt@numericable.fr