Journal of Informetrics 10 (2016) 854–862

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

From rankings to funnel plots: The question of accounting for uncertainty when assessing university research performance

Giovanni Abramo [a,*], Andrea Ciriaco D'Angelo [b], Leonardo Grilli [c]

[a] Laboratory for Studies of Research and Technology Transfer, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Via dei Taurini 19, 00185 Roma, Italy
[b] University of Rome 'Tor Vergata' and Institute for System Analysis and Computer Science-National Research Council of Italy, Dipartimento di Ingegneria dell'Impresa, Università degli Studi di Roma 'Tor Vergata', Via del Politecnico 1, 00133 Roma, Italy
[c] University of Florence, Dipartimento di Statistica, Informatica, Applicazioni 'G. Parenti', Università degli Studi di Firenze, Viale Morgagni 59, 50134 Firenze, Italy

ARTICLE INFO

Article history:
Received 23 March 2016
Accepted 7 July 2016

Keywords:
Bibliometrics
Universities
FSS
Productivity
Funnel plot
Italy

ABSTRACT

The work applies the funnel plot methodology to measure and visualize uncertainty in the research performance of Italian universities in the science disciplines. The performance assessment is carried out at both discipline and overall university level. The findings reveal that for most universities the citation-based indicator used gives insufficient statistical evidence to infer that their research productivity is inferior or superior to the average. This general observation is one that we could indeed expect in a higher education system that is essentially non-competitive. The question is whether the introduction of uncertainty in performance reporting, while technically sound, could weaken institutional motivation to work towards continuous improvement.

1. Introduction

In the current knowledge-based economy, research and higher education systems play a significant role in supporting the competitiveness and socio-economic growth of nations, through the education of white collar workers and production of new knowledge. Improvement in the research and higher education infrastructure has with good reason become a policy priority for a growing number of governments. Among the interventions to increase the effectiveness and efficiency of research institutions, and the socio-economic returns from public spending on R&D, a growing number of countries are launching national assessment exercises of their research institutions. The exercises are intended to accomplish various aims, as selected by the national governments: for informing selective funding of research institutions; stimulating better research performance; reducing information asymmetry between supply and demand in the market for knowledge; informing policy formulation and strategic decisions; and last but not least, demonstrating that public investment in research is effective and delivers public benefits. An international comparative analysis of performance-based research funding (Hicks, 2012), indicates that subsequent to the example of the original UK research assessment exercise (the RAE, in 1986), at least 14 other countries (China, Australia, New Zealand and 11 in the EU) have chosen to implement national assessment exercises as the basis for directing at least some portion of public financing for research institutions. Alongside this, several annual world university rankings continuously receive great media attention, influencing opinion and practical choices. The various

* Corresponding author.
E-mail addresses: giovanni.abramo@uniroma2.it (G. Abramo), dangelo@dii.uniroma2.it (A.C. D'Angelo), grilli@disia.unifi.it (L. Grilli).

http://dx.doi.org/10.1016/j.joi.2016.07.005
1751-1577/

national and international assessments employ a variety of indicators and methods (bibliometric, peer-review, informed peer-review, surveys) to assess institutions' research performance. A common feature of the vast majority of the performance assessments is the lack of confidence intervals indicating the likely range of population values. The scores observed for research performance are in fact related to the assumptions and limits of the particular measurement method and indicators, and in the case of aggregate measures to the different sizes of the research institutions. Accounting for the uncertainty embedded in the measurements is crucial to establish whether the performance of an institution is truly outstanding or the result of random fluctuations.

Assessment of research performance is affected by both bias and uncertainty factors.[1] Bias factors generate fluctuations with systematic effects. A typical source of bias is the differing intensity of publication and citation across fields, which the evaluator ideally attempts to limit through a fine-grained classification of scientists and field-normalization of citations. Conversely, uncertainty factors randomly affect the assessment, meaning they will generate fluctuations without systematic effects in favor or against particular groups. Typical factors increasing uncertainty in performance assessment are the variability in intensity of production due to personal events, or due to the patterns characteristic of research projects, or the varying lengths of review and publication time across journals. Ideally, uncertainty factors should again be limited, but they cannot be completely eliminated. Notwithstanding uncertainty, the analyses of research performance can still be valid, as long as the reporting includes measures of uncertainty. The recent Leiden manifesto (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015) wisely recommends that practitioners 'avoid misplaced concreteness and false precision' in reporting performance values, and that 'if uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values'.

However, indications of uncertainty are generally not provided for the popular international 'league tables' of universities. This is true whether the performance scores and relevant rankings are produced by 'non-bibliometricians', such as the Shanghai Jiao Tong University Ranking (SJTU-Shanghai Jiao Tong University, 2016), QS World University Rankings (QS-Quacquarelli Symonds, 2016) and Times Higher Education World University Rankings (THE-Times Higher Education, 2016), or whether they are produced by bibliometricians themselves, such as the Scimago Institutions Ranking (Scimago, 2016). The CWTS Leiden Rankings few years ago integrated stability intervals (Waltman et al., 2012). In our studies concerning Italian university research productivity rankings (e.g. Abramo, Cicero, & D'Angelo, 2011; Abramo, D'Angelo, & Di Costa, 2011), we have ourselves not usually provided the likely range of performance values. Recently we attempted to deal with this shortcoming, by applying a funnel plot methodology to measure and visualize uncertainty in the research performance of the institutions. The funnel plot shows the uncertainty in data values by adding confidence bands, indicating the range where research performance indicator's values are expected to lie on the basis of the institution's size. To illustrate the funnel plot methodology, we applied it to measure uncertainty in the research productivity of Italian universities active in Biochemistry (Abramo, D'Angelo, & Grilli, 2015). The results showed that just one university out of 42 had truly outstanding research performance, while for 79% of universities the performance was not different from the overall mean, at a 5% significance level. Should the results in Biochemistry be confirmed for all sciences, then any performance rankings neglecting uncertainty would be misleading for policy and decision-making.

Considering this question and its implications, the current work thus extends the application of funnel plots to measure uncertainty in all fields and disciplines of the sciences. The goal is to identify the proportion of outstanding universities in each single field of research (192 fields), namely units whose difference from the overall mean is statistically significant. The analyses will be carried out separately for each discipline (nine disciplines).

The provision of reliable institutional research performance scores, including visualization of uncertainty levels, has implications for both stakeholders and policy makers. The stakeholders can include anyone who draws on or is influenced by the rankings, from the casual observer to the potential student, to the interested enterprise and the highest political levels. Suffice it to think of the many countries that allocate public funding according to the rankings stemming from national research assessment exercises. Or how in the Italian case, parliament recently considered a proposal to normalize the graduation scores of candidates competing for public positions, by the 'quality' score of their degree-granting university.

We refer the reader to our previous work (Abramo, D'Angelo, & Grilli, 2015) for an overview of the quite limited literature on measuring uncertainty in research performance, as well as a description of the funnel plot methodology. We would like to add here a work by Claassen (2015), which was published meanwhile. The author measures uncertainty in university quality estimates by eight different world ranking systems, showing that the difference between universities ranked 50th and 100th, and 100th and 250th, is not significant.

The funnel plot methodology presents advantages over other methods for visualizing uncertainty. For example, in the popular caterpillar plot the performance assessment of the units are plotted in increasing order and endowed with confidence intervals (see Spiegelhalter, 2005; and the references therein). The lengths of the intervals summarize the uncertainty and a unit whose interval is above (below) zero is judged to have a performance significantly above (below) the overall mean. Even if a caterpillar plot is technically correct, it may not be effective in communicating the results because: (i) it does not explicitly show the relationship between the level of uncertainty and the volume or size of the units, and (ii) it leads the reader towards undue emphasis on the ranking of the units though the reliability of the ranking is not assessed (the exact

[1] We refer the reader to our previous work (Abramo et al., 2015) for a detailed analysis of all factors of uncertainty and bias in research performance assessment.