International Journal of Multidisciplinary and Current Educational Research (IJMCER)
ISSN: 2581-7027 ||Volume|| 2 ||Issue|| 5 ||Pages 108-112||2020||

Toward Predictive Modelling for the Quality Assessment of Research Outputs in the Research Excellence Framework

Mario Gianni
University of Plymouth, Plymouth, UK

ABSTRACT*:* In this work we will analyse the main strengths and weaknesses of an alternative method for estimating the starred quality levels of the research outputs in the Research Excellence Framework (REF). We will also discuss the main advantages and disadvantages of this method against the peer review process currently in place and we will conclude highlighting how far the model is from being used in practice by the assessors in future exercises.

I. INTRODUCTION AND LITTERATURE REVIEW

The Research Excellence Framework (REF) is a process of expert review undertaken approximately every six years by the four higher education funding bodies in the UK (Research England, the Scottish Funding Council, the Higher Education Funding Council for Wales, and the Department for the Economy, Northern Ireland) to assess the quality of the research across all the disciplines of the Higher Educational Institutions (HEIs). It was used for the first time in 2014, replacing the Research Assessment Exercise (RAE) which was last conducted in 2008. Similar frameworks in other countries are, for example, the AERES in France and the ANVUR in Italy which was implemented on the main structure of the REF [26].

The main purpose of this exercise is threefold: (1) to inform the selective allocation of the grants given to the HEIs by the main four funding bodies in the UK, (2) to account for public investment in research and produce evidence of the benefits of this investment, and, (3) to provide benchmarking information and establish reputational yardsticks, for use within the HE sector and for public information [3-6, 9-11]. The REF has also been proved to be beneficial for providing a rich evidence base to inform strategic decisions about national research priorities, for creating a strong performance incentive for HEIs and individual researchers and for informing decisions on resource allocation by individual HEIs and other bodies [7, 8].The REF currently classifies all the disciplines in the HE in 34 Units of Assessment (UoA) and groups them in 4 Main Panels[1]. According to this classification, the assessment of the quality of the research is carried by 34 sub-panels (one for each UoA) working under the leadership and guidance of four main panels. The sub-panels are responsible of assessing three distinct indicators of each submission which are outputs, impact and environment. In this work we will focus our attention on the criteria and procedures used by the expert sub-panels to evaluate the quality of the research outputs. We refer the reader to [9-11] for more details about the assessment of the overall quality profile of HEI.

The standard metrics used by the expert sub-panels to award the quality of the research outputs of the HEI are the so-called starred quality levels [5]. These metrics are divided in five distinct levels: four star, three star, two star, one star and unclassified. These levels are given to the research outputs depending on their originality, significance and rigour, with reference to international research quality standards [5]. The *Panel criteria and working methods* document of the REF [5] provides the expert sub-panels of each UoA with a descriptive account of how to interpret and apply the criteria for assessing outputs and the starred quality levels. Based on this description, sub-panels mark four star those research outputs whose quality is considered world leading in terms of originality, significance and rigour, and that provide evidence of research that (i) is leading or is at the forefront of the research area, (ii) develops fundamental new concepts for research and (iii) introduces major changes in practice. Thee star is given to research outputs which are internationally excellent in terms of originality, significance and rigour, but which falls short of the highest standards of excellence. Evidence of this level of quality is given by the presence in the outputs of important contributions to the field in terms of knowledge and techniques which are likely to have a lasting influence but are not necessarily leading to fundamental new concepts. The sub-panels award two star those research outputs which are recognised at international level, provide useful knowledge and influence in the field and involve incremental advances including new knowledge which conforms with existing ideas and paradigms. One star is given to the research works that are visible at national level and have minor influence in the field. Finally, by exclusion, outputs

[1] REF2021

which do not satisfy the above criteria are considered unclassified in the evaluation process. Moreover, in assessing work as being four, three, two, one star or unclassified, the sub-panels also look at the types of problems which have been addressed, the research methods and the theoretical principles which have been used, the empirical results, the intellectual precision, the appropriateness of the concepts, the theories and methodologies deployed within the research output, the integrity, coherence and consistency of the arguments. Besides the above guidelines, for some of the UoA (but not for all of them) the *Panel criteria and working methods* document of the REF [5] provides sub-panels with secondary criteria related to the impact factor, the publisher and the number of citations of the research outputs.

As it can be observed, the criteria and the starred quality levels are very general and are open to varying modes of applications. The first issue concerns with the use of citation metrics. The latest assessment REF exercise took place in 2014. It was followed by a detailed report, known as the *Metric Tide* report [12], that critically examined the possible role of citation metrics in the REF. It concluded that "[m]etrics should support, not supplant, expert judgement" [12]. To support this conclusion, the report provided statistical evidence of the lack of agreement between metrics and peer review. Interestingly, there are conflicting viewpoints on the degree of the correlation between citation metrics and peer review. Some, such as Mryglod *et al.* [13] and Mahdi, D'Este, and Neely [14], argued that a correlation of 0.7 is too low to consider using metrics, while others, such as Thomas and Watkins [15] and Taylor [16], argued that a correlation of 0.7 is sufficiently high. This indicates that different researchers draw different conclusions, despite finding similar correlations. One problem is that none of the correlations are assessed against the same yardstick; thus, it is unclear when a correlation should be considered *high* and when it should be considered *low* [18]. As a conclusion, the next exercise, planned for 2021, will also be conducted via peer review, partly because of the UK academia's continued opposition to an increased role for mechanical methods of evaluation of research output, even when several other countries do adopt a bibliometric evaluation, as highlighted in Wang, Vuolanto, and Muhonen's survey [18]. Furthermore, the use of indicators may lead to strategic behaviour and gaming [21].

Another relevant issue is related to the different forms of unconscious bias which could affect the judgement of the peer reviewers [19]. Watermeyer and Hedgecoe conducted an interesting analysis of a simulated impact evaluation exercise populated by approximately 90 senior academic peer reviewers and user assessors, undertaken within one UK research-intensive university prior to and in preparation of its submission to REF2014 [20]. Although this study was conducted on an indicator which is not considered in this work, it revealed that in their efforts to evaluate impact, peer reviewers were indirectly promoting a kind of *impact mercantilism*, where case studies that best sold impact were those rewarded with the highest evaluative scores, thus indicating that unconscious biases can really constitutes a source of concern.

In August 2018, we have been appointed by the University of Winchester as external assessors for the Mock REF2 - Main Panel C, Unit of Assessment 17 – REF2021 where we undertook a review and provided comprehensive feedback on 4 journal articles. In this personal experience, among other findings, we identified that one of the main forms of unconscious bias was related to the difficulty of being *impartial* especially in evaluating the quality of theoretical concepts and methodologies that were proposed in the assessed research outputs and that were dealing with problems that we also addressed, though in a different perspective [22]. We also interviewed some of the staff members in our department who were appointed by our institution to undertake the internal mock REF exercise in the past years. To the question regarding the main criteria that they applied to rank the research outputs, most of them agreed on the presence of special keywords in the text, such as for example, the words "novel" and "preliminary".

PROPOSED ALTERNATIVE MODEL FOR ASSESSING QUALITY: The proposed alternative model for estimating the starred quality levels of the research outputs aims to leverage the issues previously discussed by learning correspondences between the citation metrics, the assessment criteria, the research outputs and the starred quality levels.The model is composed of three main building blocks. The first block is responsible of mapping a research output into a finite set of keywords associated with quantitative values specifying the frequency of the keywords in the text [23]. The second block takes as input the research output and the related citation metrics and augments the former with a quantitative representation of the spatial distribution of the citations across the globe. This provides an estimate of the level of visibility of the research output. The last block is a classifier based on Cost-Sensitive Support Vector Machines (SVMs) [24]. It takes as input the features associated with the keywords; the citations metrics augmented with the spatial features and returns as output a label representing the estimation of the starred quality level of the research output.

We used a dataset composed of 100 research outputs evaluated in the Main Panel B, Unit of Assessment 12 - Engineering in the past REF2014 to train the parameters of the model via k-cross validation [25]. We tested the prediction accuracy on the trained model on two different datasets. The first dataset was composed of 20 research outputs whose subjects were similar to the ones in the dataset used during training. The second dataset was including 20 research outputs belonging to the same UoA of the outputs used for training but in completely different subjects. While on the first dataset we obtained an accuracy of 72.3% in correct classification, on the second dataset the accuracy dropped down to 38.3%.These preliminary results demonstrated that variations of the subjects of the research outputs in the same UoA significantly affect the performance of the estimation. This is due to the choice of the keywords used to map the text into numerical feature vectors. The keywords have the main role of encoding the types of characteristics which distinguish the different starred quality levels as it is described in the *Panel criteria and working methods* document of the REF [5]. They aim to map into single words complex qualitative concepts like *research at the forefront of*, *contribution to knowledge* and *robustness of theory*. They also encode the latent human factors of the judgement. Although these keywords were identified from the material collected during the interviews with colleagues with experience in REF exercises, they resulted to be limited to a particular sub-set of subjects in the UoA under consideration and relatively simple to model complex qualitative concepts and human behaviours.

Another critical issue in the predictive model regards the augmentation of the features with the spatial distribution of the citations. While this approach mitigates the problem of self-referencing, it does not fully represent the degree of visibility (worldwide, international and national) of the research outputs. For example, the quality of an output which has been cited by researchers in United States, India, Czech Republic and Malaysia not necessarily must be considered world leading in terms of originality, significance and rigour. Unfortunately, in the *Panel criteria and working methods* document of the REF [5] it is unclear the relation that there exists between rigour and visibility and, to the best of our knowledge, limited information is given on the argument.

Finally, the accuracy we obtained on the first dataset suffers from the limited number of samples in the dataset used for training the parameters of the model. Unfortunately, databases containing all the research outputs submitted by HEIs in the past REF evaluations with the corresponding starred quality levels are not available/accessible online yet. Results of the REF are provided to the institutions only in the form of percentages of the submissions meeting the standards for each quality level. The dataset used for training the parameters of the model did not contain enough samples to make the predictive model general enough. Moreover, the samples have been manually labelled. This process introduced a significant amount of noise in the data which limited the accuracy of the classification.

II. DISCUSSION AND CONCLUSION

The proposed model for predicting the starred quality level of the research outputs requires several improvements in order to be effectively used by HEIs prior to or in preparation of submissions to the REF. The first crucial problem to be solved regards the need of more labelled data. Unfortunately, this data could not be easily available online. Moreover, the data produced by the internal assessors could be considered confidential and then not shareable. To leverage this problem, the research outputs internally evaluated should be made public together with the corresponding ranks to all the staff members of the institution. Crucial attention must be paid to the condition dictated by the General Data Protection Regulation (GDPR) policy for sharing such data.

Moreover, the keywords encoding the types of characteristics of the research outputs should be learned as well rather than fixed a priori. This would enhance the correlation with the characteristics. Other approaches can also be adopted to ground the characteristics into semantic rules. These rules would filter the research outputs based on the starred quality levels.With the above refinements the generalisation capability and performance of the predicative model could be significantly improved. The revised version could have a huge impact on the current peer-review process under different fronts. It could be used to partially replace the process thus reducing the overall costs of the assessment. It could be adopted by expert sub-panel to further refine the ranks and it could serve as a tool for predicting the trends of research which will attract more funding in the nearby future.

Word count: ~2400.

REFERENCES

[1] RAE 2008. (July 2005). Consultation on assessment panels' draft criteria and working methods. [Online]. Available: https://www.rae.ac.uk/pubs/2005/04/docs/consult.pdf

[2] RAE 2008. (June 2005). Guidance on submissions. [Online]. Available: https://www.rae.ac.uk/pubs/2005/03/rae0305.pdf

[3] REF 2014. (July 2011). Assessment framework and guidance on submissions. [Online]. Available: https://www.ref.ac.uk/2014/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

[4] REF 2014. (January 2012). Panel criteria and working methods [Online]. Available: https://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf

[5] REF 2021. (February 2019). Panel criteria and working methods. [Online]. Available: https://www.ref.ac.uk/media/1084/ref-2019_02-panel-criteria-and-working-methods.pdf

[6] REF 2021. (January 2019). Guidance on submissions. [Online]. Available: https://www.ref.ac.uk/media/1092/c-users-daislha-desktop-ref-documents-final-guidance-for-live-site-ref-2019_01-guidance-on-submissions.pdf

[7] Building on Success and Learning from Experience. (July 2016). An Independent Review of the Research Excellence Framework. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/541338/ind-16-9-ref-stern-review.pdf

[8] Technopolis Group. (October 2018). Review of the Research Excellence Framework. Evidence Report. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/768162/research-excellence-framework-review-evidence-report.pdf

[9] REF 2021. (January 2019). Guidance on submissions. [Online]. Available: https://www.ref.ac.uk/media/1092/ref_guidance_on_submissions.pdf

[10] REF 2021. (January 2019). Panel criteria and working methods. [Online]. Available: https://www.ref.ac.uk/media/1084/ref-2019_2-panel-criteria-and-working-methods.pdf

[11] REF 2021. (January 2019). Guidance on codes of practice. [Online]. Available: https://www.ref.ac.uk/media/1086/ref-2019_03-guidance-on-codes-of-practice.pdf

[12] Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., and Johnson, B., “Metric Tide: report of the independent review of the role of metrics in research assessment and management,” Technical Report. Higher Education Funding Council for England, 2015.

[13] Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2014) “Predicting results of the Research Excellence Framework using departmental h-index: revisited,” *Scientometrics* 104, pp. 1013–1017. DOI:10.1007/s11192-015-1567-9

[14] Mahdi, S., D’Este, P., and Neely, A., “Are they good predictors of RAE scores?” Technical Report Advanced Institute of Management Research 2008.

[15] Thomas, P. R. and Watkins, D. S. (1998) Institutional research rankings via bibliometric analysis and direct peer review: a comparative case study with policy implications. *Scientometrics* 41, pp. 335 - 355 (1998).

[16] Taylor, J. (2011), The Assessment of Research Quality in UK Universities: Peer Review or Metrics?. *British Journal of Management*, 22: 202-217. doi:10.1111/j.1467-8551.2010.00722.x

[17] Traag, V. A. Waltman, L. *Systematic analysis of agreement between metrics and peer review in the UK REF*. CoRR, 2018.

[18] Wang, L. Vuolanto, P. Muhonen, R. (2014) Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives. *TaSTI Working Papers*. http://urn.fi/URN:ISBN:978-951-44-9660-8 .

[19] Leathwood, C., and B. Read. 2012. Final Report: Assessing the Impact of Developments in Research Policy for Research on Higher Education: An Exploratory Study. *Society for Research into Higher Education*. [Online]. Available: https://srhe.ac.uk/downloads/Leathwood_Read_Final_Report_16_July_2012.pdf

[20] Watermeyer, R. and Hedgecoe, A. (2016). Selling ‘impact’: peer-reviewer projections of what is needed and what counts in REF impact case studies. A retrospective analysis. *Journal of Education Policy* 31 (5), pp. 651-665. 10.1080/02680939.2016.1170885

[21] Sivertsen, Gunnar (2018) Why has no other European country adopted the Research Excellence Framework? Impact of Social Sciences Blog (16 Jan 2018).

[22] Murphy, T. and Sage, D. (2014) Perceptions of the UK’s Research Excellence Framework 2014: a media analysis, *Journal of Higher Education Policy and Management*, 36:6, 603-615, DOI: 10.1080/1360080X.2014.957890

[23] Elkan, C. “Clustering Documents with an Exponential-family Approximation of the Dirichlet Compound Multinomial Distribution”. *Proceedings of the 23rd International Conference on Machine Learning*. pp. 289 – 296. 2006.

[24] Iranmehr, A. Masnadi-Shirazi, H. Vasconcelos, N. Cost-sensitive support vector machines Neurocomputing, 2019, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.11.099.

[25] James, G. Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning*. Springer, 2013.

[26] Geuna, A. and Martin, B. R. (2003) University Research Evaluation and Funding: An International Comparison. *Minerva*. 41(4) pp. 277 – 234. DOI: 10.1023/B:MINE.0000005155.70870.bd.