Citations and Sub-Area Bias in the UK Research Assessment Process

Alan Dix
Talis, Birmingham, UK
and University of Birmingham, Birmingham, UK
http://alandix.com/ref2014/

ABSTRACT

This paper presents a citation-based analysis of selected results of REF2014, the periodic UK research assessment process. Data for the Computer Science and Informatics sub-panel includes ACM topic sub-area information, allowing a level of analysis hitherto impossible. While every effort is made during the REF process to be fair, the results suggest systematic latent bias may have emerged between sub-areas. Furthermore this may have had a systematic effect benefiting some institutions relative to others, and potentially also introducing gender bias. Metric-based analysis could in future be used as part of the human-assessment process to uncover and help eradicate latent bias.

Categories and Subject Descriptors

CCS2012: Information systems → Information systems applications → Digital libraries and archives; Human-centered computing → Collaborative and social computing

General Terms

Measurement.

Keywords

REF, research assessment,, bibliometrics, research funding

1. INTRODUCTION

At the end of 2014 the UK completed its latest round of sexennial research assessment, REF2014, the Research Excellence Framework, when more than a thousand academics and other experts divided into 36 subject panels assessed nearly two hundred thousand research outputs and other evidence provided by over 150 universities and university-level bodies [7, 8]. The evidence provided included broad statements about academic environment in 'Units of Assessment' (UoA a REF term, roughly corresponding to an academic department) and evidence of the non-academic impact of work. However, this paper focuses on the 'outputs', specific items of research output, most commonly in computer science an individual conference or journal paper.

There has been on-going discussion over the potential for metrics-based evaluation, not least because of the cost of the REF process, but this has been overwhelmingly rejected by the community (e.g. [2]). So, while some subject sub-panels within REF, including computing, have citation data available, others do not, and all take this as at best suggestive or contributing as part of broader professional judgement.

Several authors have provided post hoc analysis of previous research assessment exercises, showing broad correlations between metric-based measures and the overall grades of departments [1, 3, 4, 5, 12]. There is also broad agreement that at a suitably large level of aggregation citation-based metrics provide a useful validation or check; indeed HEFCE are using them to help ensure that differences between subject sub-panels are defensible. This paper therefore assumes that citation-metrics can be used as a valid measure of quality between *large enough* units in computing.

The REF process works on an edge between transparency and openness about process, whilst preserving the confidentiality of individual assessments of outputs. This is because its purpose is the assessment of UoAs not individuals, and that the level of effort and precision needed for the latter to be reliable is far greater than the former.

The need, inter alia, for confidentiality means that detailed output-by-output scores are destroyed soon after the assessment processes is complete. This means that, hitherto, the only knowledge of outcome from the process were the average, and more recently profile, scores for each UoA. On the other hand, the desire for transparency means that the unscored outputs, with all supporting data, is available in the public domain, including, for some panels, Scopus citation data for each output. In the Computer Science and Informatics sub-panel, this data is particularly rich allowing a level of post-hoc analysis that has not previously been possible.

The results are interesting from a bibliometric point of view and also disturbing in what they reveal about latent bias within a process that strives the utmost for fairness.

Note: The author was a member of REF 2014 sub-panel 11 "Computer Science and Informatics", and as such is bound by confidentiality. The analysis presented here is therefore based solely on public domain data and processes.

2. RICH DATA IN COMPUTING

With the exception of a small number of confidential outputs, the vast majority of information about submissions is available in the public domain from the REF2014 web site [9]. This includes the title and venue of the output, other identifying information, including DOI, ISSN, or ISBN depending on the nature of output, and crucially, where this was made available to the panel, the Scopus citation count at a fixed census date late in 2013.

Furthermore, the data available for the Computer Science and Informatics sub-panel is richer again as submitting UoAs were asked to provide a precise topic for each output based on the ACM taxonomy of computing sub-areas.

During the REF process, each output was awarded a level from 4* (world-leading) to 1* (recognised nationally). However, as noted, the actual level score for each output is not public domain and will be destroyed entirely. Only the overall score profile for each UoA

is provided giving a percentage of outputs in each level, which is then used as part of funding formulae or decisions. For most panels this is the only knowledge of the scores.

For computing, however, we unusually have additional public domain information. Before the level scores were destroyed Morris Sloman produced a number of statistical analyses based on these, the Scopus citation data and also Google citation data [11]. The last was not available during the assessment process, but collected for the post-hoc analyses. Critically, one of Slomans' results was a 4*/3*/2*/1* profile for each of the ACM sub-areas (excerpts in figure 1), which is widely available and reproduced in the REF Panel B final report [10].

Topics	% Rating within Topics				
	4	3	2	1	GPA
Cryptography	45.5%	38.2%	10.9%	5.5%	3.2
Real-time and fault-tolerant systems	40.9%	31.8%	22.7%	4.5%	3.1
Logic	33.4%	50.5%	16.1%	0.0%	3.2
Computer vision	33.2%	45.2%	19.3%	2.3%	3.1
Algorithms / Theory / Methodologies	32.2%	48.6%	16.3%	2.9%	3.1
Computer graphics	27.8%	43.9%	25.9%	2.4%	3.0
Models of computation / formal languages /	27.3%	51.4%	20.7%	0.7%	3.1
Networks (protocols)	14.0%	52.9%	27.3%	5.8%	2.8
Modeling and simulation	13.8%	50.0%	33.0%	3.2%	2.7
Human-centered computing / Visualization	10.0%	48.9%	34.3%	6.7%	2.6
Collaborative and social computing	8.8%	46.9%	36.3%	8.1%	2.6
Other Topics: OR, History, Education etc	5.9%	31.4%	44.1%	18.6%	2.2
Applied computing: law, humanities, educat	5.1%	38.1%	43.2%	13.6%	2.3
Total	22.1%	47.2%	25.7%	4.7%	

Figure 1. Sub-area REF 2014 profiles (excerpts from [10,11]).

3. INTERPRETING SUB-AREA DATA

The REF process was aimed at evaluating and comparing complete units of assessment (UoAs), not sub-areas of computing. The computing sub-panel report therefore adds the following warning to the data released:

> *"These data should be treated with circumspection as they represent a single snapshot of outputs selected just for REF2014 and were gathered primarily to help in the allocation of outputs, where they were very useful."* [10, p.48].

However, the data do show a very clear trend. If ranked by number of 4* outputs, more theoretical areas top the table with over 30% 4* while more applied areas tend towards the bottom with often below 15% 4*. Despite the warnings, many universities are already using this chart as an indication of areas on which to concentrate looking forward to REF2020.

Ignoring for a moment the practical implications for the discipline, this additional information provides an opportunity for more detailed post-hoc analysis than has hitherto been possible for UK research assessment results.

4. METHODS

Multiple forms of citation data were available for each output:

(i) *Raw Scopus citation data*. This was either blank, meaning no citation data available in the Scopus data, or an integer including 0 (data available and zero); empty citation data was regarded as a missing value, not zero. In order to make this comparable across years, quartile levels were calculated for each year and each output given a quartile score from q1 (lowest 25%) to q4 (top 25%). Overall 22% of papers were awarded a 4*, so the top quartile corresponds to roughly the same number of outputs (but not the same outputs) as 4*.

(ii) *Normalised Scopus citation data* was created using a 'contextual data' table provided by the REF team from Scopus data [6]. The table gives typical citation patterns for different sub-areas of computing. This allows each output to be assigned a world-ranking within its own sub-area, correcting for the way that some sub-area tend to have higher citation counts than others. The ranks were 1% (best, among top 1% of outputs globally), 5%, 10% and 25%.

(iii) *Google scholar citations*, usually assumed to be more reliable for computing publications. These were initially drawn from the same scrape used in Sloman's analysis, but with an additional verification step (as the matching had been automatic). Just under 200 citations were hand corrected as they linked to the wrong Google scholar record. While the number was small statistically, it seemed prudent to remove even this level of potential noise. Google largely knows about things because they are cited, so there is no equivalent to present but zero. Analyses were therefore repeated with zero treated either as 'present but zero' or 'missing value'; however this never made any substantial difference to results. Like the Scopus data, this was reduced to quartiles within years to give cross-year comparable results.

Both (i) and (iii) analyses were repeated for all the years and also restricting to 2008-2011 only as most 2012/13 outputs have few citations.. For (ii) the normalising data was only available for 2008–2011 anyway. This led to seven separate analyses: (1) Scopus all years, (2) Scopus 2008–2011, (3) normalised 2008–2011, (4) Google all years with no citations as missing value, (5) as (4) for 2008–2011, (6) as (4) but with no citations treated as present and zero, (7) as (6) for 2008–2011.

While results were more or less extreme for each case they gave the same overall story. Where particular graphs or data are presented they are typical or conservative, not cherry-picked.

For each sub-area of computing the variants of the citation analysis end up with either a quartile score, or a top n% score (for normalised citations). To give a common base these can each be converted into predicted 4*/3*/2*/1* levels and these compared with the actual profile's from Sloman's analysis [11]. Some subject areas were small so, while figures were computed for these also, they are not used when making comparisons below.

5. RESULTS

While the details differed between the various analyses, the pattern was similar: more formal/theoretical areas (e.g. logic, algorithms) tended to have 2-3 times more 4* ratings than predicted from citations, whereas more applied and human–centric areas between 2/3 and 1/3 the predicted level. Other classic, but more mixed areas, such as AI and Vision were close to predicted figures. Figure 2 shows just how little relationship there is between citations and REF scores. The vertical axis ranks the subject areas by number of 4* outputs, the horizontal axis by number of outputs in top quartile of citations. As is evident there is no discernable correlation or pattern.

At first this appears to contradict earlier work that has suggested a strong correlation between citations and ratings in previous UK research assessment exercises [1, 3, 4, 5, 12]. However, these were focused on institutional rankings (the only data available). Indeed, Sloman's analysis [11] showed an overall correlation between citation and REF ratings on an output-by-output basis, and in *some* measures this is also true on an institutional comparison. The large divergence is at the level of subject areas.

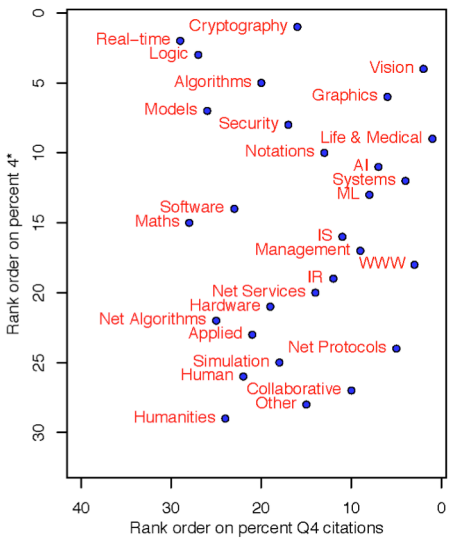

Figure 2. REF 4* vs citation ranks

6. DISCUSSION AND FURTHER WORK

Those who oppose metrics-based assessment (as the author did until this exercise) could see the divergence as validation of the need for human judgement over numbers. Furthermore, the winners and losers reflect widespread opinion within computer science as to the relative value of areas. However, the normalised analysis especially makes it hard to argue that papers ranked in top 1% of their area worldwide can be up to 10 times more or less likely to be 4* depending on the area.

Given many of the areas are large enough for reliable comparison (many hundreds of outputs), the effects are real and suggest that, despite the best efforts of the panel, latent inter-area biases have emerged.

As well as potential distortions in hiring and other strategic decisions within institutions, it is possible that these inter-area biases may have affected the overall outcomes. Above it was noted that at an institutional level *some* measures correlate well between citations and REF ratings, notably research 'power' (average rating times staff count),. However, figure 2 shows that the number of 4* is particularly sensitive and this disproportionately affects funding.

When a metric close to the value used to apportion funding is used, the citation–REF correlations become weak and exhibit systematic effects. Of 27 institutions where the REF ratings on this funding-metric fall 25% or more below the citation prediction (the 'losers'), 22 are post-1992 universities; and of the 18 that are 25% or more above (the 'winners') 17 are pre-1992 universities.

It is possible that there are other factors at work (e.g. halo effects when assessing papers from 'good' institutions), but it also seems likely that inter-area bias is partially responsible: many newer universities have greater emphasis on applied areas, and several of the biggest 'winners' are institutions with a large formal/theoretical side to their work.

In future it would be good to model the effect of sub-area bias on institutional assessment and also examine possible impact on gender bias as there are differences in male/female participation between sub-areas.

Looking towards REF2020, there were peculiarities to the algorithmic process used in the computing panel that corrected for inter-expert differences, but not overall biases. Assuming human assessment will continue to be central, providing data, such as that in this paper, within the process could help to identify and correct potential bias early in the process, hence improving the overall robustness and fairness of outcomes, which is the ultimate desire of all involved.

Links to all material used in the analysis, full analysis spreadsheets, and all raw data will be available at:

http://alandix.com/ref2014/

7. ACKNOWLEDGMENTS

Thanks to various colleagues who have discussed aspects of this work over the past few months and particularly Andrew Howes who independently replicated the analysis and provided figure 2.

8. REFERENCES

[1] Clerides, S., Pashardes, P. and Polycarpou, A. (2011) 'Peer review vs metric-based assessment: testing for bias in the RAE ratings of UK economics departments', *Economica*, vol. 78(311), pp. 565–83. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=982219

[2] Lipsett , A. (2007). Institute criticises proposed RAE replacement. *The Guardian,* Thursday 13 December 2007. http://www.theguardian.com/education/2007/dec/13/research assessmentexercise.highereducation

[3] Oppenheim, C. (1995) The correlation between citation counts and the 1992 Research Assessment Exercises ratings for British library and information science departments, *Journal of Documentation*, 51:18-27.

[4] [Oppenheim, C. (1998) The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology, *Journal of Documentation*, 53:477-87. http://dois.mimas.ac.uk/DoIS/data/Articles/julkokltny:1998:v:54:i:5:p:477-487.html

[5] Oppenheim, C. and Summers, M. (2008). Citation counts and the Research Assessment Exercise, part VI: Unit of assessment 67 (music). *Information Research*, 13(2), June 2008. http://www.informationr.net/ir/13-2/paper342.html

[6] *REF contextual data*, 2013 (spreadsheet)

[7] *Research Excellence Framework 2014*. http://www.ref.ac.uk/

[8] Research Excellence Framework 2014 (2015). *Output profiles and diversity*. http://www.ref.ac.uk/results/ analysis/outputprofilesanddiversity/

[9] Research Excellence Framework 2014 (2015b). *Results and submissions*. http://results.ref.ac.uk

[10] Research Excellence Framework 2014: *Overview report by Main Panel B and Sub-panels 7 to 15*, January 2015. http://www.ref.ac.uk/panels/paneloverviewreports/ panelminutes/

[11] Sloman, M. (2015). *Sub-panel 11 Computer-Science and Informatics REF Analysis*.

[12] Smith, A., and Eysenck, M. (2002) "The correlation between RAE ratings and citation counts in psychology," June 2002 http://psyserver.pc.rhbnc.ac.uk/citations.pdf