

The CWTS Leiden data Analysis of the 2018 release

The data made available in the now annual ‘CWTS Leiden Rankings’ (<http://www.leidenranking.com/>) could potentially provide useful insights into the performance of the UK research base. This analysis uses that data as a proof of principle, showing what might be learned from them. It is not a formal part of any UKRI decision-making, evaluation or assessment process.

The 2018 ranking data contains information on citations relating to 55 countries which between them host 938 universities. Forty eight of these universities are in the UK (listed in Annex 3). Publications are allocated to five ‘fields’ of research.

Previous years’ data are presented alongside the most recent release, allowing the analysis of changes over time within a consistent, openly-available and well-documented framework.

Visualising quality with Leiden data.....	2
The association between volume and quality.....	6
The changing citation landscape	11
Performance by field	12
Summary	16
Annex 1 – Description of Leiden data.....	17
Annex 2 – The relationship between publication volume and citation metrics.....	18
Annex 3 – UK HEIs appearing in 2018 Leiden data.....	19

Visualising quality with Leiden data

The key Leiden indicators are the percentages of publications which are in the top 1%, 5%, 10% and 50% of comparable publications¹. It is assumed that the more exclusive the banding, the better the publication is likely to be. The 1% range is likely to be quite noisy and will not be used in this analysis.

Starting at the highest level, the proportion of each country's publications across all fields combined which is in the top 10%, the range 10% to 50% and the bottom 50% of publications is shown as a ternary diagram in Figure 1:

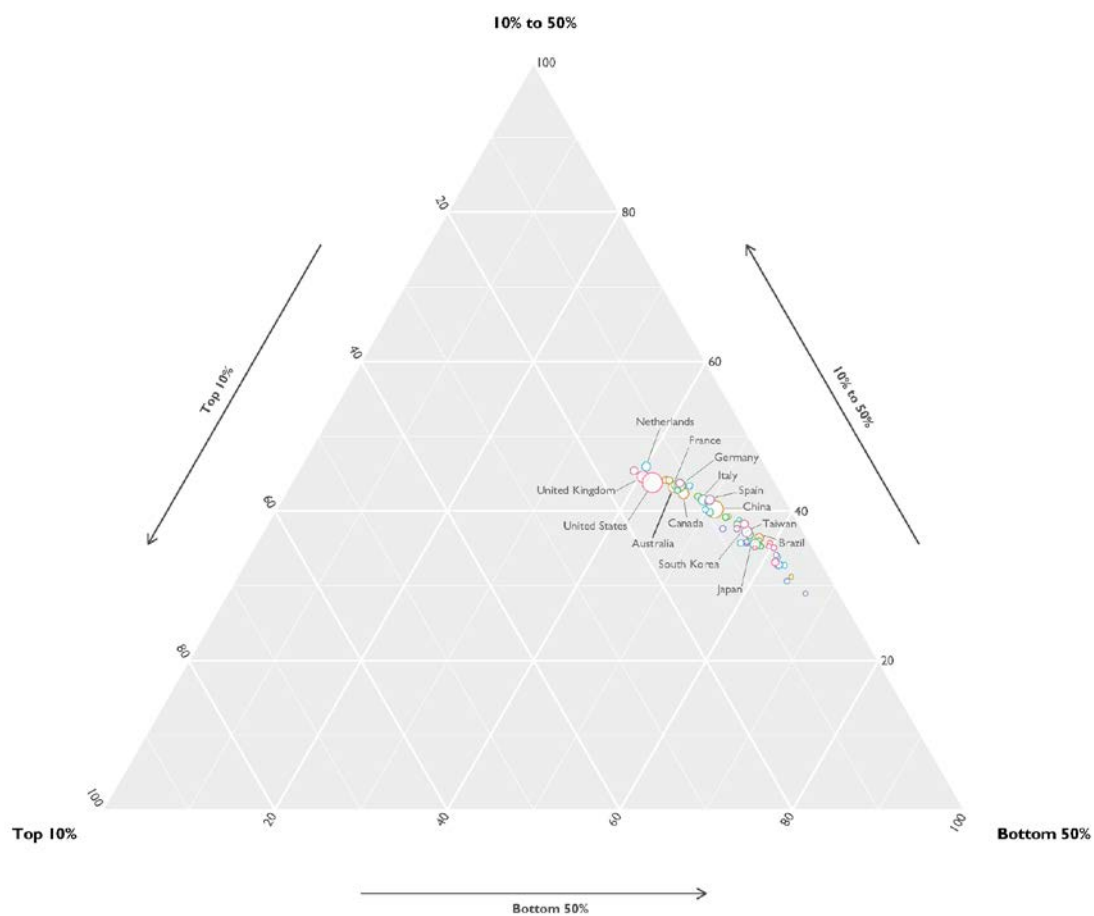


Figure 1: ternary diagram for Leiden 2018 data. All 55 countries are shown but only those with > 50,000 publications are labelled directly. Circle areas are in proportion to the number of publications in the data.

Ternary diagrams can be used to show the composition of three-component mixtures for which the total composition sums to 100% (as it does in this case by definition.) Each apex of the triangle indicates a 100% composition of that particular element – top 10%, 10% to 50% and bottom 50% – and each base 0% of the composition of the opposite apex.

Countries with less than 10% of publications in the top 10% are underperforming by that measure, as are countries with more than 50% in the bottom 50%. It is of course entirely

¹ Other indicators relate to the extent and nature of collaboration on outputs, and relative citation measures. This analysis makes use of only a fraction of the data.

possible to be under-performing by one measure and over-performing by another, and nothing can be inferred from the data about the reasons for apparent under-performance.

A sequence of countries' progress towards global scientific influence is apparent in the arc starting towards the bottom right apex and ending closer to the centre of the chart. A key feature to note is that all countries lie on this arc: there are no outliers, most of the possible space is completely unoccupied and observed behaviour quite tightly bounded. Figure 2 zooms into the area of interest².

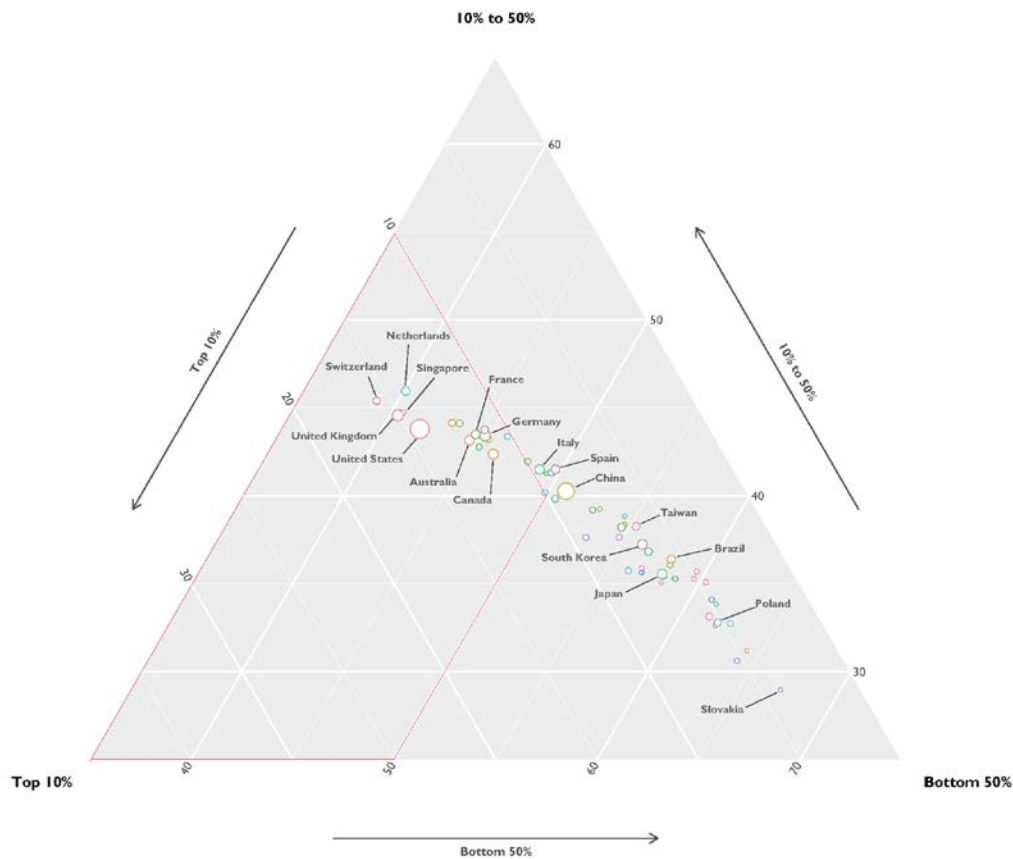


Figure 2: zoomed image of Figure 1. Additional countries labelled and trapezium of high-performing nations outlined in red.

The UK is squarely in a region of over-performance³, alongside accepted comparator nations such as the US, Switzerland and Germany. If placement on the arc is understood qualitatively as an indicator of success then the UK is performing well and is ahead of most of its competitors.

Japan's placement is unexpected. A country's position reflects a mix of how good its outputs are, how good people think they are, and how likely it is that people will see them in the first place. Leiden data only index publications written in English, and it seems that language effects could be at work. In countries which do not have English as their first language a degree of filtering, which may or may not be independent of quality⁴, must occur, with not

² Subsequent charts are all zoomed for ease of presentation, but the tight clustering should be borne in mind.

³ The red trapezium in Figure 2, demarcated by the 10% and 50% lines for 'top 10%' and 'bottom 50%' outputs.

⁴ Though it is tempting to believe that the better outputs are more likely to find their way into more international (and hence English-speaking) journals as it might be worth the additional effort needed to do so.

all native language publications appearing in English. It would be unwise to conclude without further justification that Japan, South Korea or Taiwan actually do underperform.

Most of the main European nations – plus Australia, Canada and the USA – are found in the high-performing zone. Italy and Spain alone fall outside it, being found closer to China than to other European countries.

Figures 1 and 2 show the results of the 2018 data (that is, the citations recorded for outputs published in the period 2013-2016.) This static view can be augmented by looking at trends across earlier releases of Leiden data⁵.

With a total of eight sets of Leiden data it would be impractical to represent $8 \times 55 = 440$ data points simultaneously on a single chart, so the following chart focuses on just seven comparator countries: Canada, China, France, Germany, Japan, UK and the United States. Figure 3 shows each country's performance in each category in each of the eight years:

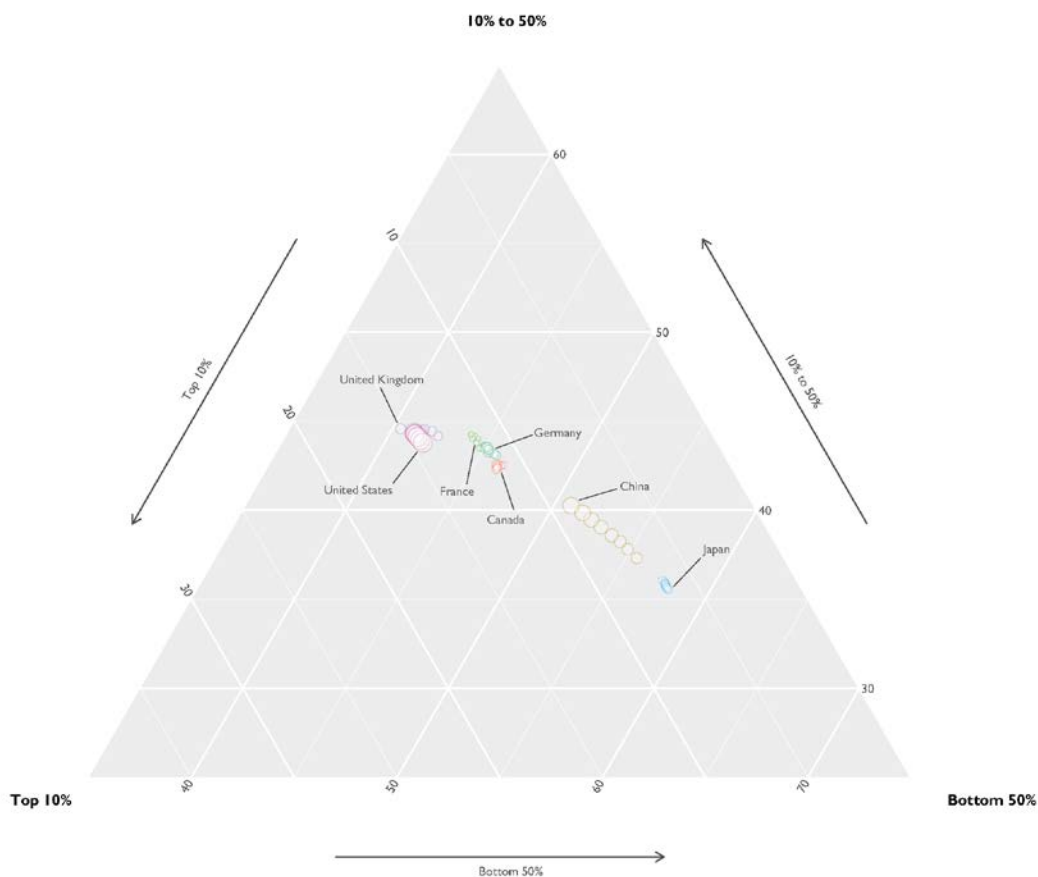


Figure 3: performance as reported in Leiden data over the eight years to 2018 for seven leading countries. Leader line touches 2018 (that is the most recent) national performance for each country. Note the zoomed axes.

China's improving trajectory and increasing scale of activity is very clear. Its general trend is of a shift away from the bottom 50%, and more towards the top 10% than the mid-range. The USA's trajectory is opposite in direction but represents only a minor change overall,

⁵ Eight periods are covered in total, each covering a three year (overlapping) window of publication dates.

Figure 5 shows the performance over time of ten leading UK HEIs:

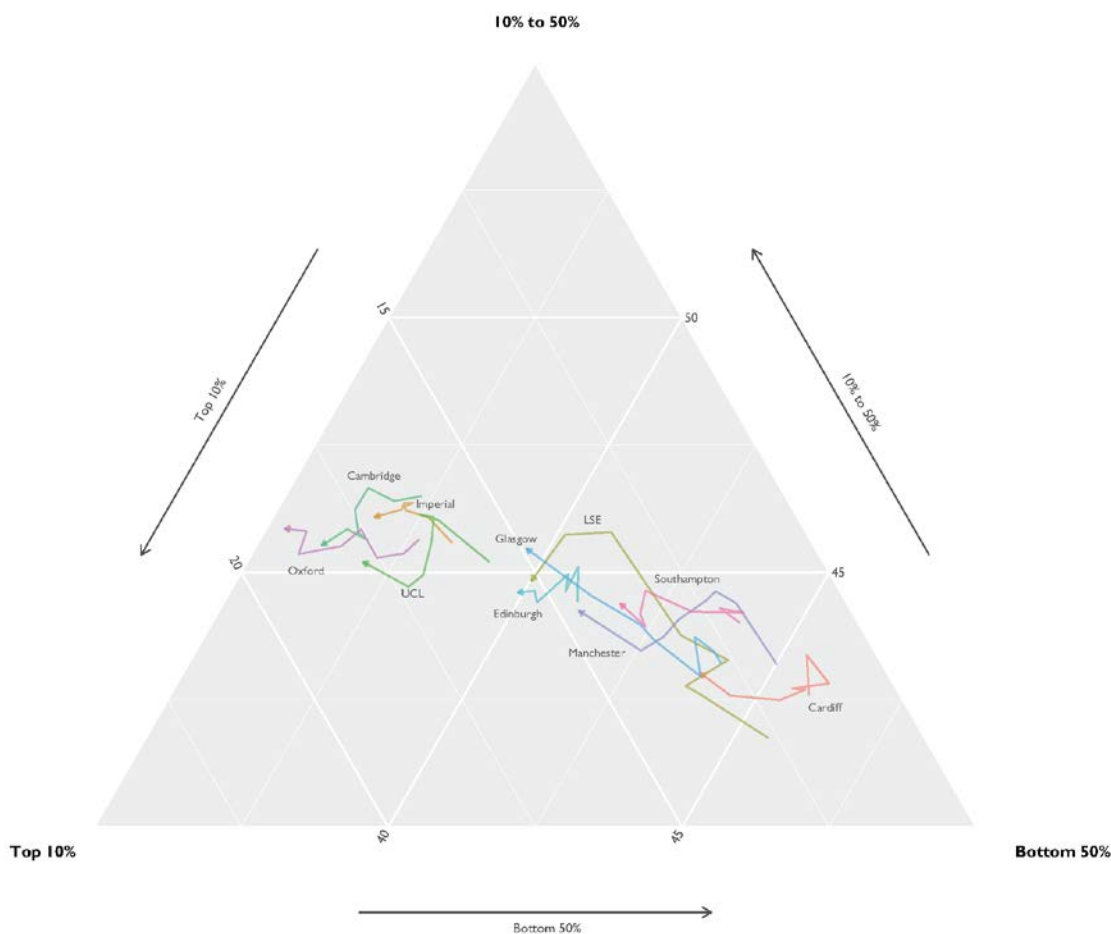


Figure 5: performance as reported in Leiden data over the seven years to 2018 for nine leading UK HEIs. Arrows show direction of travel over data period. Note the zoomed axes.

Most UK HEIs have seen a steady trend of movement away from the bottom 50% and towards the top 10%. In some cases the changes over the period are slight and/or erratic (Edinburgh, Imperial) and in some they are relatively large and consistent (Glasgow, Manchester.) LSE's performance is particularly notable in that they have reduced a tail of outputs in the bottom 50% by more than 5%: the largest directional change of all these organisations.

Overall the UK's direction of travel, individually and collectively and as indicated by Leiden data, is noticeable and positive. Again though it must be emphasised that there may be more than one cause for the change, and that it would not be safe to assume that everything we see is necessarily a sign of actual improved performance⁷.

The association between volume and quality

⁷ For example, without further analysis a shift of resources away from lower-performing fields would be indistinguishable from a general improvement across all fields. See Figure 11 for that analysis.

As already suggested in Figure 2, but by inspection only, there is an association between volume of outputs and their relative quality (5%, 10% and 50% levels.) In this section this observation is put on a firmer footing⁸.

Figure 6 shows the proportion of each country's outputs which is in the top 5%, 10% and 50% against the total volume of outputs in the 2018 Leiden data and, as a dashed line, the calculated expected proportion of outputs in each category.

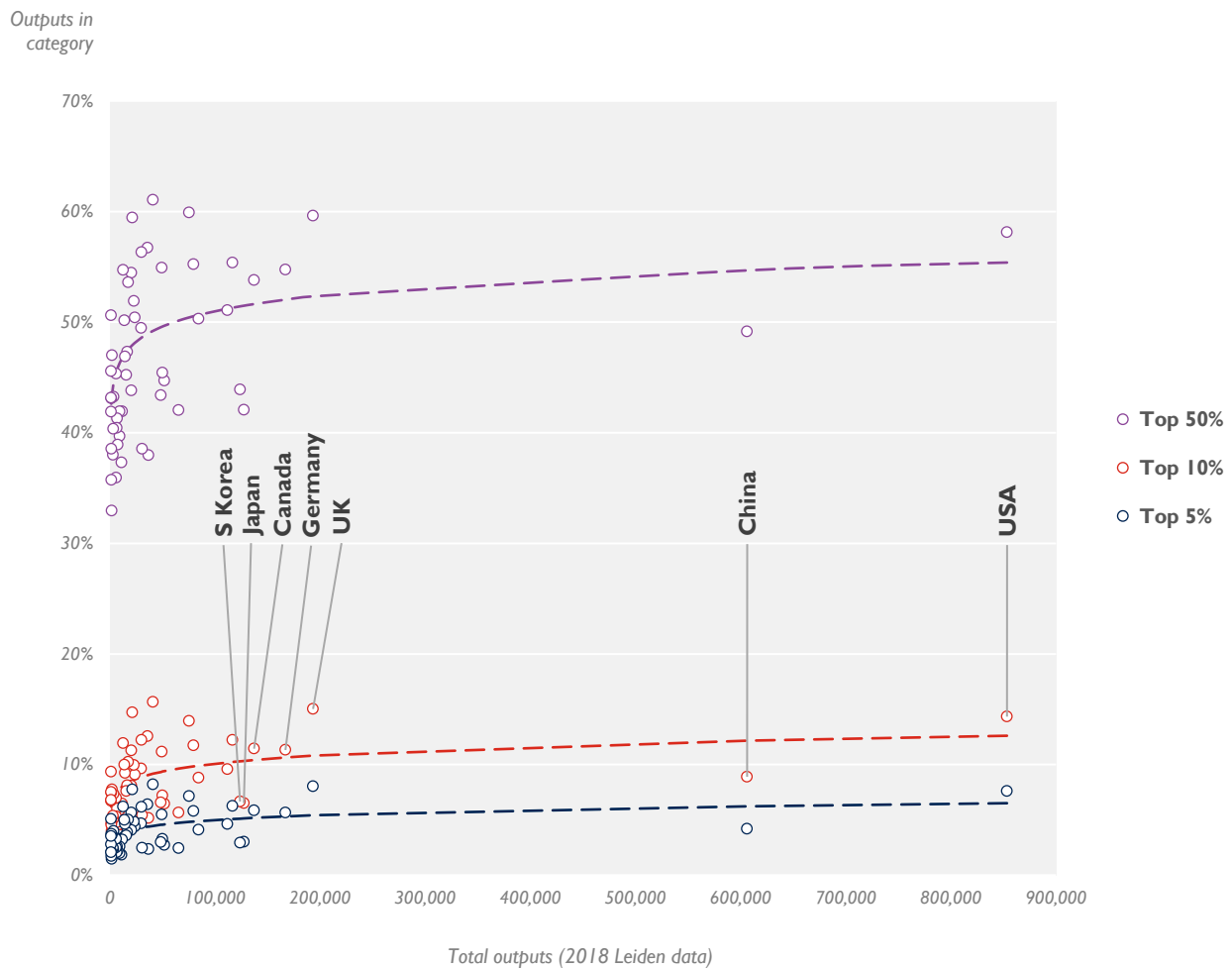


Figure 6: association between total output volume and % of outputs in top 5%, 10% and 50% of all publications for Leiden data 2018. In all cases the association (plotted as a dashed line) is significant at below the $p = .05$ level.

The relative under-performance of China, Japan and South Korea in all three categories, unexpected given their publication volumes, is again apparent⁹.

If we assume that a country (or HEI) performing as expected would have 5% of its publications in the top 5%, 10% in the top 10% and 50% in the top 50% then, based on the

⁸ Based on the approach outlined in 'The relationship between decision volume and success rates' (<https://esrc.ukri.org/about-us/performance-information/application-and-award-data/>) and also see Annex 2

⁹ With such large numbers of publications confidence intervals for estimated proportions are too small to be worth showing.

relationships established in Figure 6, it is possible to calculate the minimum number of outputs which is required to match the expected proportion of outputs in each category¹⁰ (5%, 10% and 50%.) These are shown in the table below for country-level 2018 data.

Level	Required 2018 country output volume
Top 5%	103,300
Top 10%	93,200
Top 50%	61,200

As the category becomes smaller and more exclusive, the projected number of outputs required to at least meet expectations increases, from 61,200 for the top 50% range all the way to 103,300 for the top 5%. Only those countries which produce a significant volume of publications might routinely expect to find their publications in the upper echelons of research outputs.

The same general patterns are true within the UK for each HEI's performance both within years and also over the full period of the Leiden data. Taken over the whole period, UK HEIs that published more tended to be cited more (Figure 7):

¹⁰ This does not imply a causal relationship between quality and volume, it is just a description of the data (though it is also possible that such a relationship exists – we do not know.)

Outputs in category

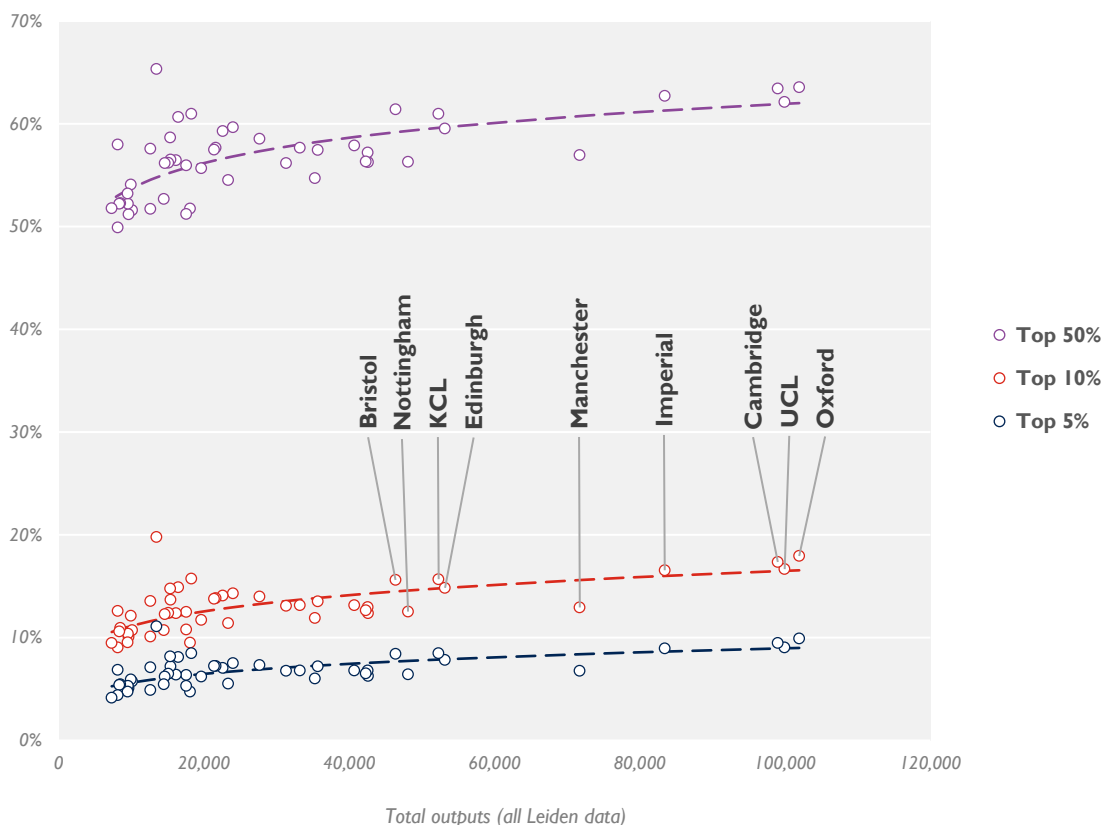


Figure 7: association between total output volume and % of outputs in top 5%, 10% and 50% of all publications for UK HEIs in the Leiden data 2011-12 to 2018. In all cases the association (plotted as a dashed line) is significant at below the $p = .05$ level.

Almost no UK organisations perform below expectations, in terms of having less than $x\%$ of their outputs in the top $x\%$ of outputs. Using the relationship established it is possible to estimate a UK performance-related volume threshold for 2018 (or indeed any period) across all publication volumes¹¹:

Level	Required 2018 HEI output volume
Top 5%	660
Top 10%	600
Top 50%	370

Based on these figures we might expect a UK university not in the Leiden data but nonetheless publishing in the 2018 period to need at least these numbers of outputs in order not to be underperforming in these terms.

¹¹ Leiden data contain no information on HEIs producing fewer than 1,000 publications in each qualifying period

It would then be *possible* to interpret these figures as an indicator of necessary critical mass in productivity and quality within the UK research system. As a rough indicator, UK HEIs producing fewer than about 200 eligible (fractionally-counted) publications per year in the three-year period would have been unlikely to have a proportionate number of publications in the top 5% and 10% of all publications. An HEI would have to produce perhaps fewer than 120 per year to risk not having half in the top half.

A UK HEI producing such small volumes is likely to be a specialist institution which probably shouldn't be compared with its much larger competitors. While the mathematical relationship might hold, this is probably an example of a metric that should not be used. A more reasonable interpretation is that there are probably no UK institutions with a small volume of generally comparable publications which are likely to be systematically under-performing.

The changing citation landscape

Extending the line of thought that seeks to understand how many publications are needed to perform at a particular level, Figure 8 shows how the annual minimum publication volume needed if a country is to at least meet each specified performance threshold has changed across all years of the Leiden data.

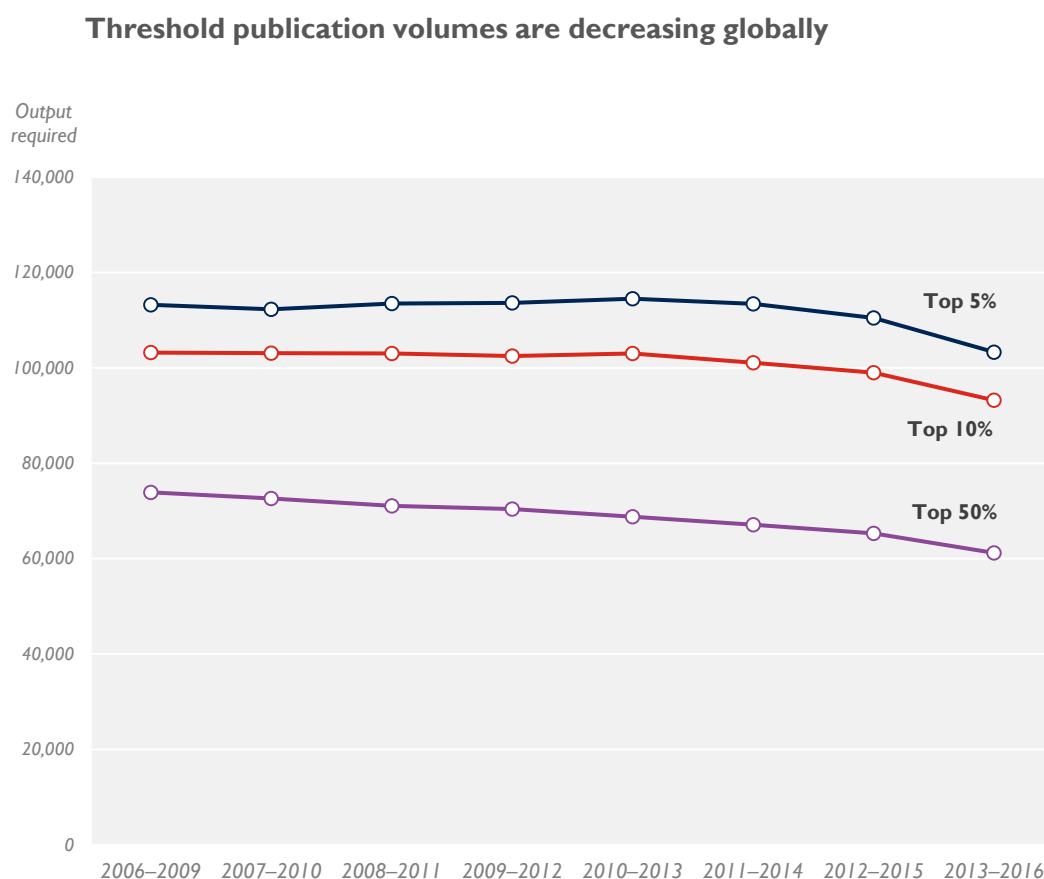


Figure 8: expected number of publications required for a country to have at least 5%, 10% and 50% of its outputs in the top 5%, 10% and 50% of all publications globally in the Leiden data 2011-12 to 2018. In all cases (year-threshold combinations) the association between volume and output quality is significant at below the $p = .05$ level.

Over the last eight years, the most highly-cited publications have become less likely to be a product of those countries which have the highest output volumes. The balance of bibliometrically-defined quality is shifting slowly away from more established, higher-volume nations, and it is increasingly untrue that ‘more active = higher quality’. But it is still very much the case that larger volumes are associated with higher citation levels and there is some way to go before the volume-citation association disappears¹².

The same general trend is found in the UK. The threshold publication volume required to be performing as expected is decreasing slowly, and less research-active HEIs are

¹² If the trend continues, with the threshold for being in the 50% zone decreasing by about 1,600 publications per year on average, it would take between 30 and 40 years for the global volume-quality association to disappear entirely, and for the more established nations to lose their volume-derived advantage.

increasingly likely to see their outputs highly-cited (Figure 9, though note the earlier caveat about the interpretation of this figure in a UK-only context):

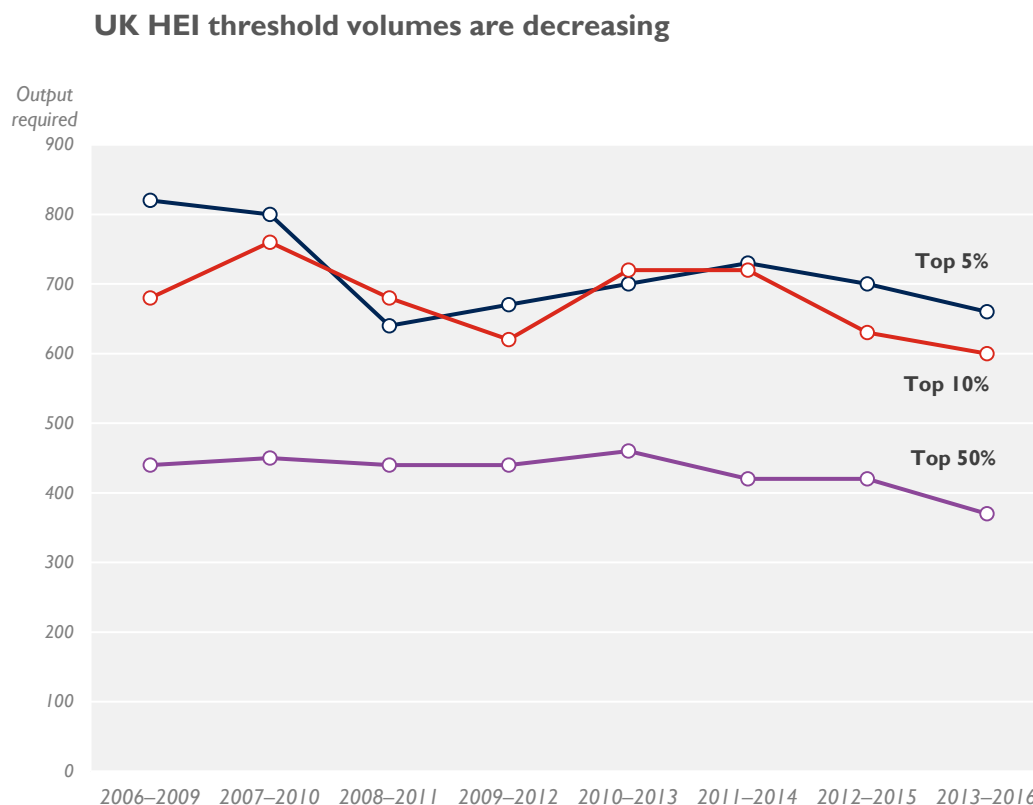


Figure 9: expected number of publications required for a UK HEI to have at least 5%, 10% and 50% of its outputs in the top 5%, 10% and 50% of all publications globally in the Leiden data 2011-12 to 2018. In all cases (years, thresholds) the association between volume and output quality is significant at below the $p = .05$ level.

However, the UK trajectory is much less smooth than for the country comparison globally¹³. This is likely to be at least partly a result of the smaller numbers and greater noise in the data. But it is also possible that the upswing in volume required to be at threshold in the period 2008-2011 to 2011-2014 may be related to the REF exercise which was based on publications produced in the period 1 January 2008 to 31 December 2013. A process of backlog clearing on the part of those able to generate a backlog in the first place may have skewed the data to reflect research performance in earlier years.

Performance by field

Leiden data are split into five 'fields':

- Biomedical and health sciences
- Life and earth sciences
- Mathematics and computer science
- Physical sciences and engineering

¹³ It is also slower. Based on the overall trend it will take around 50 years for the quality-volume advantage to disappear at the 50% level, though the rate of change does appear to have been increasing more recently.

- Social sciences and humanities

Figure 10 shows each of the 55 nations' 2018 performance in each of these fields (panels a to e) and also the UK's overall performance in each field individually (panel f).

In each case the arc of countries is apparent though with, as expected given the lower volumes, greater scatter. Non-English speaking nations are less likely to appear at the head of the arc, and non-English speaking nations that do not use the Roman alphabet are particularly disadvantaged.

In general the placement of each country is approximately the same in each field. There does not seem to be such a thing as 'a country with particular strength in...' although this does of course depend on how that categorisation is decided. Field-associated mobility along the arc is quite limited.

The relative performance of fields within countries can also be shown, but must be treated with great caution. As panel f shows, 'Life and earth sciences' appears as the UK's top-performing field, with 'Mathematics and computer science' and 'Social sciences and humanities' relatively lower. But as stated on the CWTS website:

“The Leiden Ranking does not take into account conference proceedings publications and book publications. This is an important limitation in certain research fields, especially in computer science, engineering, and the social sciences and humanities.”

Leiden data compares like with like within a field, but comparisons across fields would only be reasonable if the underlying data were comparable – we cannot assume this and we do not know how they differ¹⁴.

¹⁴ For example if all the best research in a field appears in conference proceedings, and the UK disproportionately prefers to publish in that way with journals being a last resort option but one for which the UK enjoys an inherent advantage for various reasons, then the relative performance of UK publications in that field as represented in the Leiden data will appear low. This may or may not happen in practice, but the point is that we cannot rule it out.

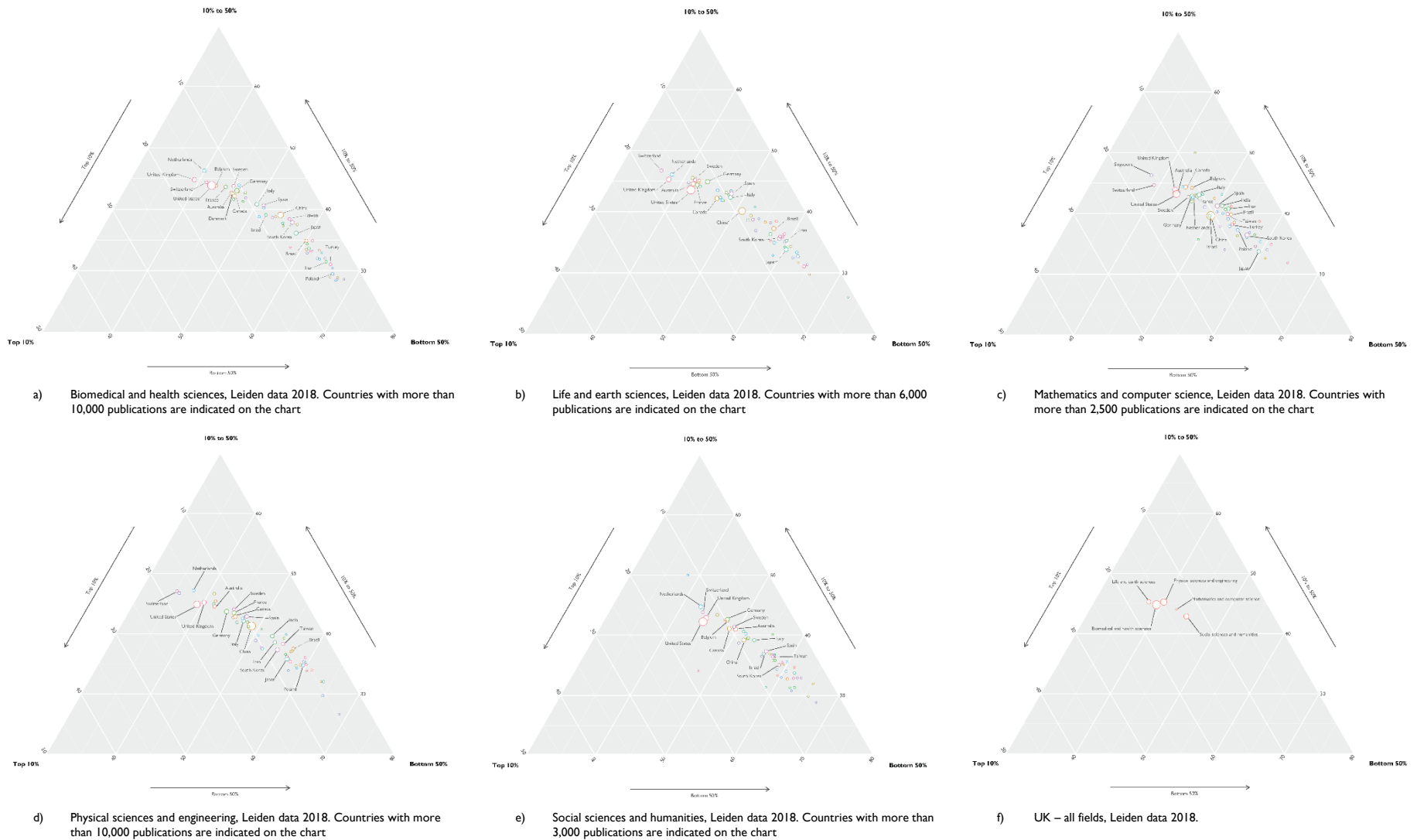


Figure 10: national performance by field, Leiden data 2018. Countries indicated are determined by threshold publication number indicated in chart. All axes are truncated.

As a final example of what might be learned from the Leiden data, Figure 11 shows how the UK's performance in each of the five fields and across 'All sciences' has changed across the data period:

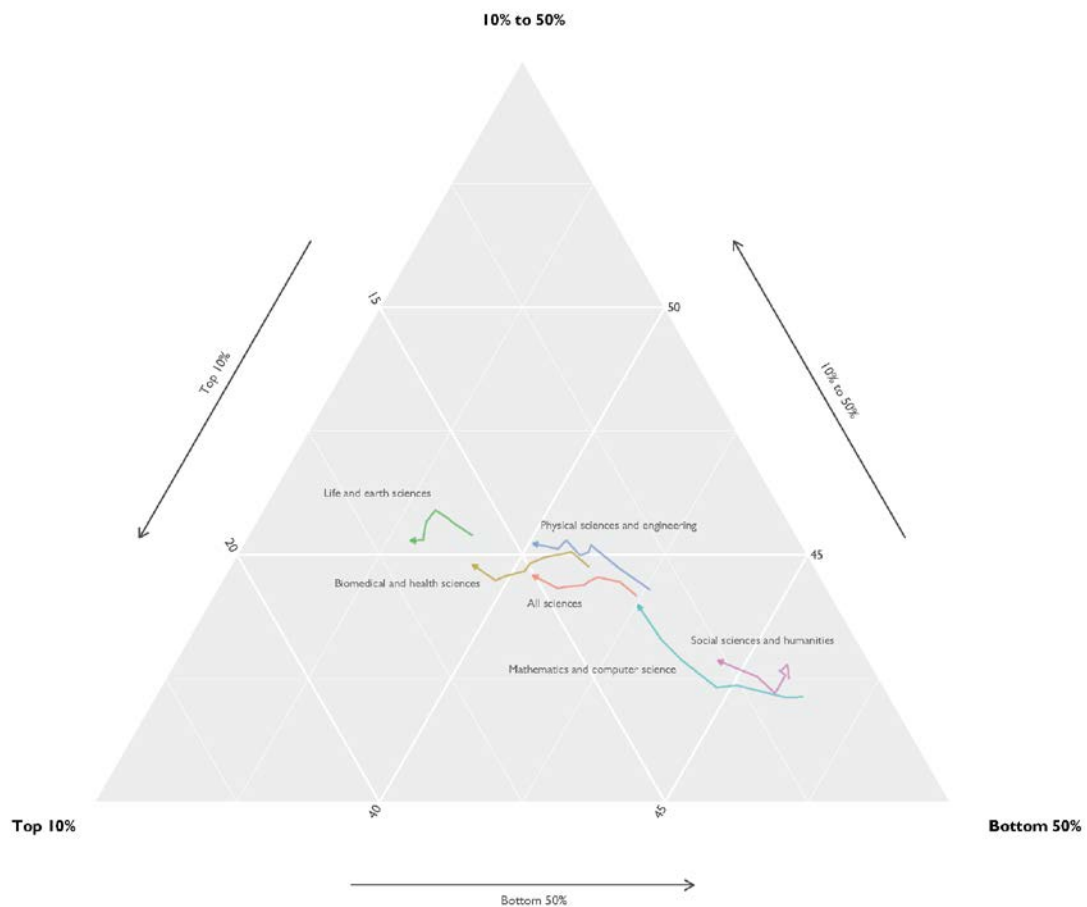


Figure 11: UK performance as reported in Leiden data over the seven years to 2018 for all Fields in which UK HEIs published. Arrows show direction of travel over data period. Note the zoomed axes.

The UK's improving performance in all fields is clear, though qualitatively different for each field. This rules out the possibility that our overall performance improvement is a result only of a shift of publications away from fields which under-perform relatively speaking. It is possible that is part of the explanation but it cannot be all of it, and the truth would only emerge with further investigation.

Summary

One thing that those who come into contact with bibliometrics on a regular basis can agree on is that bibliometrics should be used with caution, or better still ‘responsibly’¹⁵. Another thing that most of them agree on is that metrics, and in particular bibliometrics, should be usable, useful and in the end used in some form. But in terms of consensus it’s downhill all the way from there on.

Quite what form a responsible use of metrics might take no one can yet agree, and there is a lot more said about what it isn’t (*h*-indices as evidence for promotion panels, use of citation counts to award funding, journal Impact Factors as a proxy for the quality of an individual paper) than what it is.

In the absence of a strong prescription what we find is a pattern of individualised, specialist, local, bespoke, intermittent and, above all, cautious application. It’s hard to imagine UKRI not using bibliometrics in some form for some purpose(s) at some point, and the series of reports on the performance of the UK research base sets a precedent for this¹⁶. The relative merits, strengths and weaknesses of each particular instance of use of bibliometrics can and will be debated in great detail. The key issue is the extent to which they can provide useful, reliable insights into the performance of the UK research base. We can only determine this by engaging with the debate.

¹⁵http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/The,Metric,Tide/2015_metric_tide.pdf

¹⁶ <https://www.gov.uk/government/publications/performance-of-the-uk-research-base-international-comparison-2013>; <https://www.gov.uk/government/publications/performance-of-the-uk-research-base-international-comparison-2016>

Annex I – Description of Leiden data

Leiden data focus on impact and collaboration.

- ‘Impact’ means the number of citations received by a publication that appears in the Social Sciences Citation Index or the Arts & Humanities Citation Index, both of which appear in Web of Science (WoS) data. The focus is on the proportion of publications produced by an organisation that appears in the top 50%, 10% and 1% of comparable publications. 5% data have been added for 2018.
- ‘Collaboration’ refers to instances of publications included in an organisation’s set of publications which feature collaborators from organisations other than that of the author whose affiliation leads to the inclusion of that publication in that organisation’s list of publications. Collaborations with any other organisation and those with industry in particular, international collaborations and collaborations classified as close (<100km) and far (>5000km) geographically are counted and included in the data.

The inclusion of multi-author publications means that either fractional or full counting of publications can be used. In this analysis fractional counts (for example if a paper has five authors from five different organisations, 0.2 of a publication is registered with each organisation) are used for the impact indicators.

Only publications in international scientific journals which have a citation pattern suitable for analysis are included. Taken together this means that the data does not describe fully all the outputs of all disciplines. But as the data are comparable, comparisons between organisations can be made on the same basis. How useful those comparisons or their underlying indicators are is a matter of judgement.

Only organisations with at least 1000 WoS-indexed publications in each three-year period are included.

Overall the data are very specific in what they describe, but reliable in terms of those comparisons.

Annex 2 – The relationship between publication volume and citation metrics

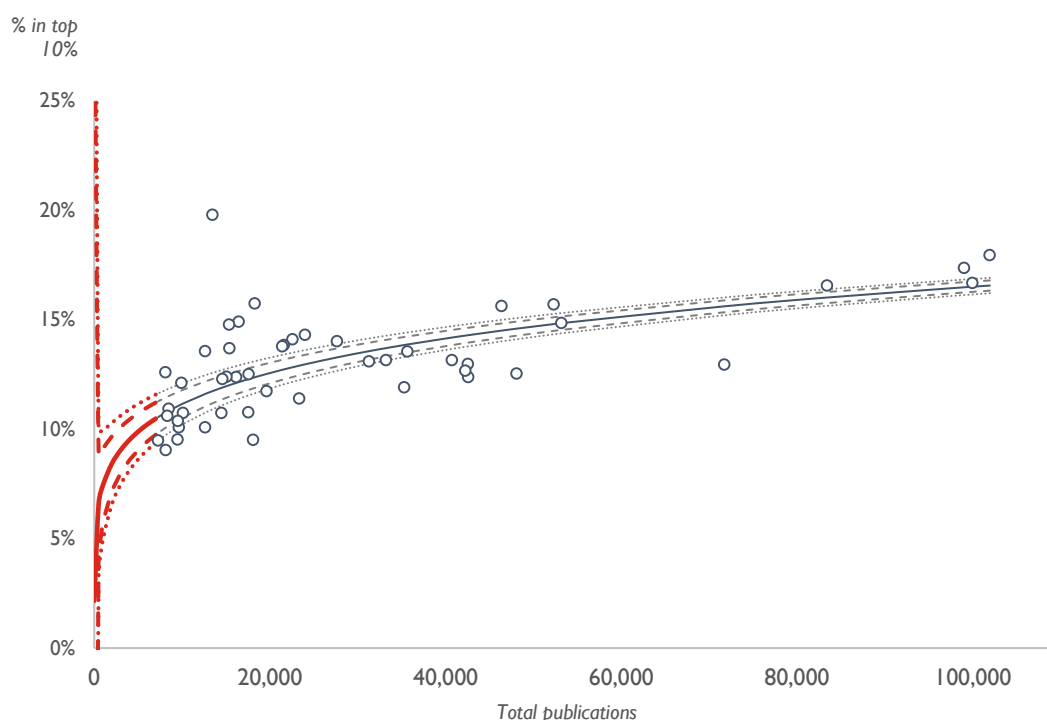
As explained in ‘The relationship between decision volume and success rates’ (<https://esrc.ukri.org/about-us/performance-information/application-and-award-data/>) an underlying relationship between success and volume, where in this case success indicates a publication that is within the top x% of publications and volume is the total number of publications produced by an HEI or country can be modelled with a function of the form:

$$\text{Success rate} = 1 - e^{-an^{(b-1)}}$$

This is bounded by 1 and 0 and takes the value 0 when the number of publications, n , is 0. When:

- $b > 1$, volumes and success are positively associated;
- $b < 1$, volumes and success are negatively associated;
- $b = 1$, $n^{(b-1)} = 1$, there is no success-volume relationship

values of a and b for each year and comparator set can be derived from the Leiden data. Once derived the relationship can then be used to calculate the expected proportion of publications within each band for HEIs or countries missing from the data due to their low publication volumes. For example, the figures for UK HEIs across all Leiden data periods:



There are no HEIs with fewer than 7,000 publications in the data, but rates can be predicted for them anyway – these are shown in red. The chart also shows the confidence limits associated with the predicted proportions, which can be used to identify HEIs with unusual performance – these are not plotted in the main Figures for clarity.

Annex 3 – UK HEIs appearing in 2018 Leiden data

Bangor University
Brunel University London
Cardiff University
City, University London
Cranfield University
Durham University
Heriot-Watt University
Imperial College London
King's College London
Lancaster University
LSE
London School of Hygiene & Tropical Medicine
Loughborough University
Newcastle University
Open University
Queen Mary University of London
Queen's University Belfast
Swansea University
University College London
University of Aberdeen
University of Bath
University of Birmingham
University of Bristol
University of Cambridge
University of Dundee
University of East Anglia
University of Cambridge
University of Dundee
University of East Anglia
University of Edinburgh
University of Exeter
University of Glasgow
University of Hull
University of Kent
University of Leeds
University of Leicester
University of Liverpool
University of Manchester
University of Nottingham
University of Oxford
University of Plymouth
University of Reading
University of Sheffield
University of Southampton
University of St Andrews
University of Strathclyde
University of Surrey

University of Sussex
University of Ulster
University of Warwick
University of York