Institute for Scientific Information

Unpacking research profiles: Moving beyond metrics

Jonathan Adams, David Pendlebury, Ross Potter and Gordon Rogers

Clarivate™

Author biographies

Jonathan Adams is Chief Scientist at the Institute for Scientific Information (ISI)™. He is also a Visiting Professor at King's College London, Policy Institute, and was awarded an Honorary D.Sc. in 2017 by the University of Exeter, for his work in higher education and research policy. ORCiD: 0000-0002-0325-4431. Web of Science ResearcherID: A-5224-2009.

David Pendlebury is Head of Research Analysis at the Institute for Scientific Information. Since 1983 he has used Web of Science™ data to study the structure and dynamics of research. He worked for many years with ISI founder Eugene Garfield. With Henry Small, David developed ISI Essential Science Indicators™. ORCiD: 0000-0001-5074-1593. Web of Science ResearcherID: C-7585-2009.

Ross Potter is a Lead Data Scientist at the Institute for Scientific Information. He has extensive research experience within academia, including NASA related postdoctoral positions at the Lunar and Planetary Institute, Houston, Texas, and Brown University, Providence, Rhode Island. ORCiD: 0000-0002-1338-5910. Web of Science ResearcherID: R-3590-2019.

Gordon Rogers is a Lead Data Scientist at the Institute for Scientific Information. He has worked in the fields of bibliometrics and data analysis for the past 10 years, supporting clients around the world in evaluating their research portfolio and strategy. ORCiD: 0000-0002-9971-2731. Web of Science ResearcherID: ABA-6554-2020.

Foundational past, visionary future

About the Institute for Scientific Information

The Institute for Scientific Information at Clarivate™ has pioneered the organization of the world's research information for more than half a century. Today it remains committed to promoting integrity in research while improving the retrieval, interpretation and utility of scientific information.

It maintains the knowledge corpus upon which the Web of Science™ index and related information and analytical content and services are built. It disseminates that knowledge externally through events, conferences and publications while conducting primary research to sustain, extend and improve the knowledge base.

For more information, please visit **www.clarivate.com/isi**

ISBN: 978-1-7395906-4-2
DOI: 10.14322/isi.grr.unpacking.research.profiles

Executive summary

This is the second report from the Institute for Scientific Information on the value of shifting from simple metrics of research activity and performance to visually more informative profiles. These profiles help us understand what is going on in research and so enable better policy and management decision making. We focus on four key indicators at researcher, journal, institutional and research field levels.

01 Introduction

This report addresses the conflict between simple metrics and deeper visual exhibits in data analysis, emphasizing the value of profiled data over metrics. We discuss the skewed distribution of research activity data, the limitations of average citation impact and raw citation counts, and the concept of normalization using Category Normalized Citation Impact (CNCI). Through the lens of four key indicators, we examine excessive self-citation, characteristics of journals, the influence of international collaboration on research performance indicators, and the use of Research Fronts™ to identify current impactful research.

02 Self-citation: what is excess? – David Pendlebury

Self-citation is a normal and expected practice by which authors relate their current work to their previous publications. But how much is too much? Should evaluators worry that the citation data used in an assessment may be anomalous, perhaps gaming research credit? To aid in validation, we explore graphical analysis of self-citation data, showing variation in cultural norms between disciplines while confirming that every discipline has a consistent central range which highlights more doubtful outliers.

03 Journal characteristics – Gordon Rogers

There are an increasingly wide range of descriptive profiles now available for the more than 21,000 journals indexed in the Web of Science, expanding the information in the annual Journal Citation Reports™ beyond the well-established Journal Impact Factor™ (JIF). In this report we review indicators of national orientation as part of continuing our exploration of new perspectives on the role, content and significance of journals, better informing researchers about the optimal venues for their papers.

04 Collaboration-CNCI – Ross Potter

A single 'average' metric obscures proper comparison of numbers of high and low cited papers in any data set. We previously showed how Impact Profiles visualize the real spread of citation impact. Another component that is often hidden is the influence of well cited internationally co-authored papers on an institutional or national average. We show how Collaborative Citation Impact (Collab-CNCI) can be deconstructed, pointing to where impact comes from and enabling better evaluation of domestic and international activity.

05 Research Fronts – Jonathan Adams

A major limitation to information acquired through bibliometrics is that analysis inevitably looks back in time: citations to prior papers about earlier projects. Resource management and policy decisions would be better if they did not rely on data about past achievements. To overcome this information 'lag' we show how information comes from looking forward or close to the edge of research: the Research Fronts that show where the cutting edge of research is located.

06 Conclusions

Data visualization is improved and made easier for the user to interpret if those users – in this case, researchers and research managers – can comment on the ease of interpreting the graphics and acquiring the information they need. We invite readers of this report to provide that feedback to us so that when these indicators are added to products it will be in an effective and user-friendly form.

01 Introduction

There is a conflict in data analysis between simple outcomes that are readily reviewed, such as single-figure metrics, and the more complex exhibits that describe the underlying activity. The default option, for time-limited research managers and policy makers, is to use the simple metrics but in doing so they may miss essential information that can aid interpretation, explain unexpected results and guide future investment.

Examples of simple metrics are researcher h-index, average citation impact and league table rank. Our previous report, *Profiles, not metrics* (Adams et al. 2019), demonstrates why profiled data are much more valuable than such metrics. We showed how each of these metrics hide the real evidence about research achievements and trajectories. We now continue that theme with four further examples of the way in which the Institute for Scientific Information (ISI) drills into and unpacks simple metrics with new, more visual analyses, to enable better interpretation of research information and use of research resources.

A key to understanding the need for profiles, not metrics, is to be aware of and understand the underlying statistical distribution of the data. Research activity is characterized by many people, projects and papers that add modestly to the advance of knowledge and a few that have a much greater academic, innovative and socio-economic impact.

This has long been recognized by such awarding bodies as the Nobel Foundation. However, for our purposes, the important consequence is that data distributions are highly skewed: many low values and a few exceptionally high values. This contrasts with what we often refer to as a 'normal' or bell-shaped distribution, symmetrical around the center.

Because we encounter many 'normal' distributions, in which the average is at or near the center and there is a balance between high and low values, we often wrongly assume that any statistic we are shown is also from a normal distribution. In a skewed distribution of research activity data this assumption is seriously misleading: the average is much higher than the median (or central point).

Only a profile of the data reveals the small number of people with relatively high income, large research groups, prolific publications and highly cited papers.

Citation counts are the common currency of academic achievement, especially in science, medicine and engineering, though rather less so in social sciences and only marginally in the humanities. Citations are seen as a first order approximation of research influence or impact (Garfield 1955) and more frequently cited work is associated with higher peer esteem in, for example, national assessment exercises. The problem is that citation rates and counts vary.

For example, the average citation impact for the United Kingdom is widely, and correctly, reported as about 1.4 compared to a world average of 1.0. It is therefore a surprise when we discover that more than half of U.K.-authored papers are cited *less* often than world average. The overall national average is high because of the small but important numbers of papers that are cited much more often than world average.

A key to understanding the need for profiles, not metrics, is to be aware of and understand the underlying statistical distribution of the data.

The key characteristics of raw citation counts are that they accumulate over time, they do so at a rate that depends on the discipline, and they are generally higher for reviews than standard journal articles. For example, a biology paper published 10 years ago will typically have more citations than a chemistry paper of the same age as well as more citations than more recent biology papers. For this reason, a statistic like the h-index, which just counts papers and their citations, is of limited value.

To create meaningful analyses of large collections of academic papers (i.e., original research articles and reviews in journals), we standardize their citation counts before aggregating them. We do this by comparing the observed citation count for each paper with the global average for all similar papers: same journal subject category, same publication year and same document type. This is called 'normalization', giving us the Category Normalized Citation Impact (CNCI) of each paper and the average CNCI for a group.

In this report we look again at four aspects of research activity and academic publishing that deliver a more rounded view.

i. We look at the **individual and their publications** and consider the question: what is excessive self-citation? This is a matter of increasing concern when suspect publications appear to be proliferating and the validity of research publication statistics is under threat.

ii. We examine **journals and their characteristics**. Beyond counts of outputs and citations, can we identify important characteristics that typify groups of serials?

iii. We consider **research performance indicators** and the **influence of international collaboration**. How can we account for the citation boost collaboration appears to give and display the outcome in a format that enables ready interpretation and management response?

iv. We use **Research Fronts** to look to the future. Most bibliometric data are retrospective analyses since they draw on citations to prior papers that describe even earlier work. Can we get closer to the current Research Front and identify the topics, institutions and researchers making a mark now?

02 Self-citation: what is excess?
– David Pendlebury

There is nothing intrinsically objectionable to self-citation: it is a normal and expected practice by which authors who are focused on a specific topic relate current work to their previous publications. But how much is too much? And when should evaluators worry that the citation data used in an assessment may be anomalous or even an attempt to game the ecosystem of research credit, prestige and reward? The integrity of the research system is threatened by those who depart from established cultural expectations and norms.

Citations in the scientific and scholarly literature link related content – topics, ideas and methods – and serve as indicators of research influence and impact. Citation counts and metrics have become important in the context of researcher evaluation – to guide and inform decision-making for appointments, promotions and funding. As with all scientometric indicators, which are proxies for direct observation, the gathering and interpretation of citation data is subject to confounding phenomena, such as variation in citation rates by research field and publication age. Self-citation, too, can be a confounding aspect of citation-based evaluation if the level is unusually elevated.

The use of relative or normalized citation indicators is foundational to responsible evaluation in association with peer review. If only raw counts are employed, individuals in fields with high rates of citation, such as molecular biology and genetics, would nearly always have superior scores to those working in fields with modest rates of citation, such as mathematics or computer sciences. Senior investigators typically claim much higher citation totals than junior investigators who have relatively few and more recent publications. Adjusting for these differences, as well as other things such as article types and even co-authorship (as we explain in section 4), is not only meaningful but necessary.

Similarly, self-citation rates can be calculated for a set of publications of individuals and groups. Since there is no 'standard' for self-citation across fields, normalization to find a typical pattern in each field must first be obtained. ISI analysts examine typical and expected rates of self-citation each year while compiling and analyzing data for the Highly Cited Researchers program. This annual list identifies individual researchers who have demonstrated significant and broad research influence in their field(s). This designation has become more than an academic accolade – it is a powerful currency benefiting both individuals, who are then often promoted or recruited, and institutions, which benefit in rank and reputation from having many Highly Cited Researchers on staff. It is an honor which comes with personal rewards as well as peer and industry respect which often leads to career opportunities. Academic and research institutions place emphasis on their tally of Highly Cited Researchers for promotional opportunities, and as a result there is a great desire and occasionally pressure on researchers to be included.

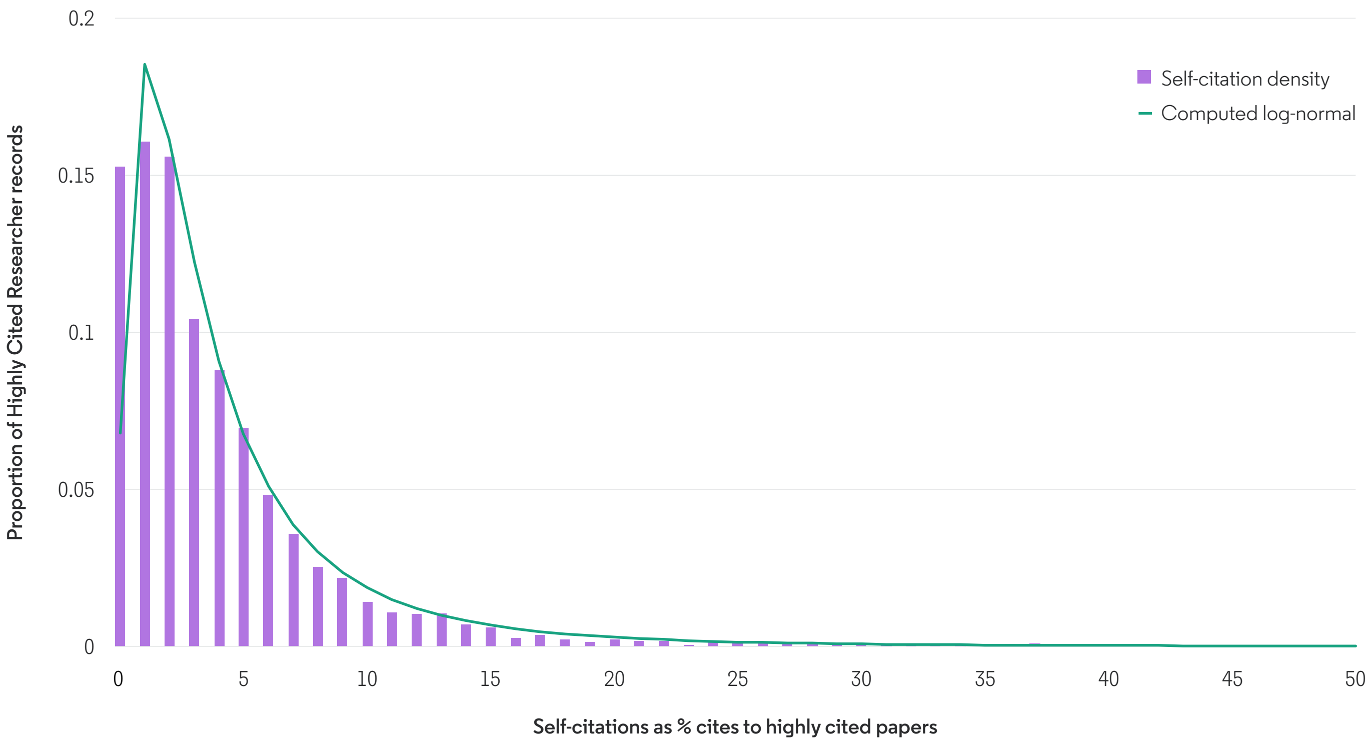
> **Highly cited is defined as papers that rank by citations in the top 1% for their field and year of publication.**

In recent years, ISI analysts have detected and filtered out individuals who engage in excessive self-citation designed to achieve Highly Cited Researcher status. Szomszor et al. (2020) describes one approach for this analysis, which uses a graphical, distribution-driven assessment of indicative excessive self-citation that demarcates a threshold not dependent on statistical tests or percentiles (since for some fields all values are within a central 'normal' range).

Highly cited is defined as papers that rank by citations in the top 1% for their field and year of publication. Figure 1 shows the distribution of the average percentage of self-citations from highly cited papers for Highly Cited Researchers, including a computed log-normal fit to a negative binomial distribution.

Evidently, self-citation generally accounts for much less than 10% of the citations received for these papers. We should therefore generally expect it to be very low. Since citation rates vary between fields, we should also expect different rates of self-citation by field.

Figure 1: The frequency distribution of self-citation (2008-2018) among Highly Cited Researchers. Numbers are shown as a proportion of 3,517 Highly Cited Researchers. A log-normal distribution is also plotted for these data. [Figure 1 (b) in Szomszor et al. (2020)].



The differences in self-citation rates between fields are indeed substantial, confirming the need to consider self-citation in the right context.

As a general guide, we can group data at the level of the 21 journal categories in Essential Science Indicators™ (ESI), which represent major domains such as Clinical Medicine and Engineering. Observation shows that most fields follow a broadly similar pattern: a few exceptionally low self-citers, most researchers in the mid-range – perhaps suggesting a cultural norm for that field – and a few high outliers. The key parameters that describe the central range in the distribution of self-citation rates are a good starting point: these are the median (mid-point), the lower quartile (LQ) and upper quartile (UQ). This confirms that typical self-citation is indeed field-dependent (Table 1).

To set an informative threshold that might signal self-citation markedly outside field norms, we use the value for the upper quartile (UQ) and the value for the inter-quartile range (IQR = UQ – LQ). If we add one, two or more IQR values to the UQ then we are setting benchmarks stretching progressively out into the realms of relative excess. A value more than 1.5 times the IQR value beyond the UQ is, statistically speaking, an outlier.

The differences in self-citation rates between fields are indeed substantial. Interestingly, fields such as Biochemistry have high innate citation rates but low self-citation whereas Mathematics and Engineering have higher self-citation associated with lower expected rates. These confirm the need to consider self-citation in the right context: high self-citation in Mathematics is a typical response to a broad subject with many small and specialist fields of research, not a sign of egregious behavior.

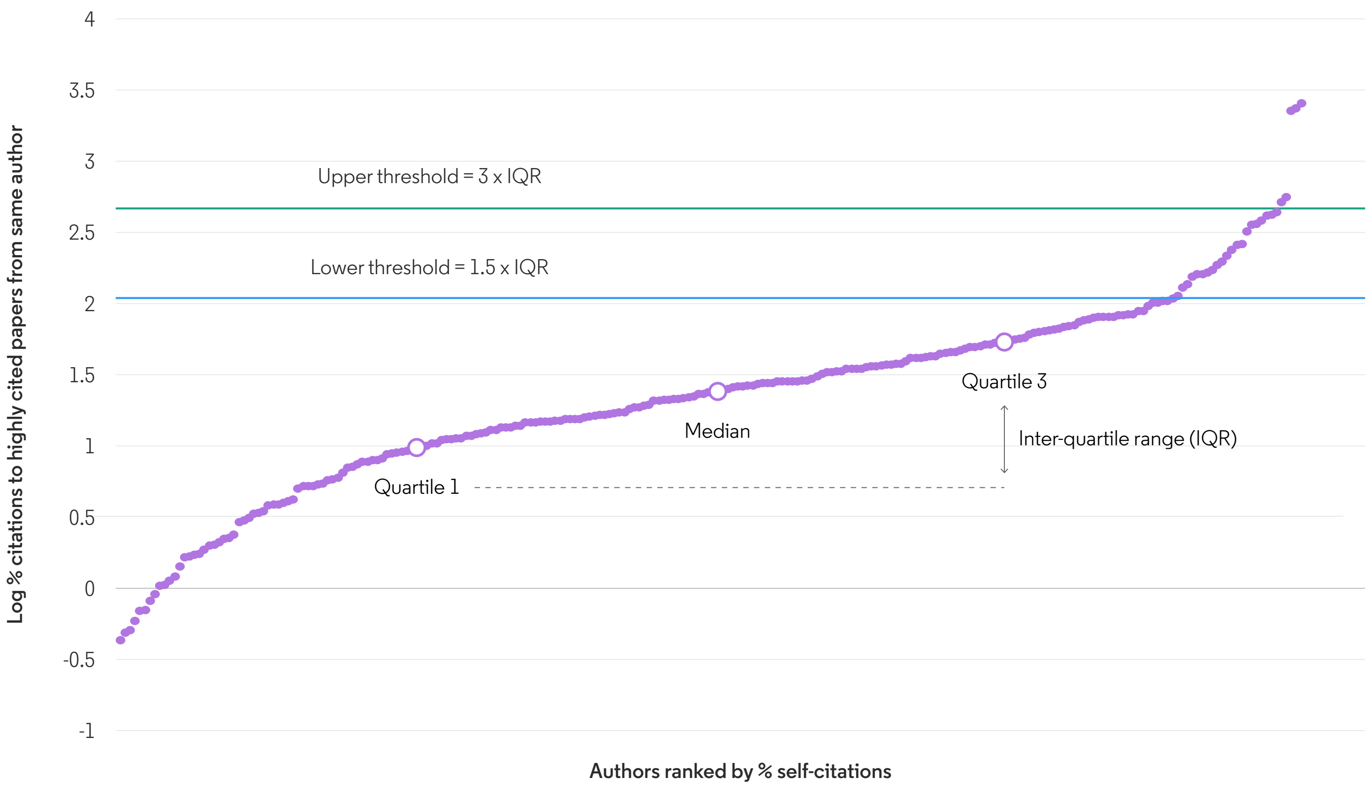
Table 1: Key parameters for percentage self-citations for highly cited papers (2008-2018), grouped by six of the Essential Science Indicators (ESI) categories. Fields are ranked by median percentage of self-citation. The inter-quartile range (IQR) is the difference between the lower and upper quartiles. The benchmark of UQ + 1.5 IQR statistically defines outliers.

ESI field	Lower quartile	Median	Upper quartile	Benchmark: UQ + 1.5 IQR	Benchmark: UQ + 3.0 IQR
Biology & Biochemistry	0.25	0.62	2.15	4.99	7.84
Physics	2.12	3.06	4.78	8.77	12.77
Psychiatry & Psychology	1.92	4.43	6.46	13.27	20.09
Agricultural Sciences	3.13	5.20	9.75	19.68	29.61
Engineering	3.14	6.01	11.01	22.81	34.62
Mathematics	6.02	12.91	20.61	42.48	64.36

These benchmarks can be used to set a percentage of author self-citation above which an individual would be eliminated from consideration for the designation of Highly Cited Researcher, even when geared to specific fields and median rates of self-citation. The distributions for the highly cited papers of candidate individuals in each field are a guide for outliers. Plotted data show how the IQR becomes a rapid indicator of the degree of departure from typical behavior (Figure 2).

A graphical display, whether linear or log, merely draws attention to data that *may* be outside behavioral norms but these depictions cannot themselves determine outliers as evidence of gaming. As ever, the data demand deeper exploration combined with expert interpretation, but at least the number of cases for analytical scrutiny may be reduced substantially.

Figure 2: Log plot of the self-citation distribution for the ESI field of Chemistry. The plot illustrates the key parameters discussed in the text: the lower and upper quartiles of the distribution and the median percentage of self cites. The lower and upper thresholds of 'typical self citation rates' are set at 1.5 and 3 times the Inter-quartile Range (IQR) and highlight statistically indicative outliers.



03 Journal characteristics – Gordon Rogers

In *Profiles, not metrics* we drew attention to the increasingly wide range of descriptive profiles that are available for all journals indexed in the Web of Science, and which expand the information in the well-established Journal Impact Factor (JIF) as part of the annual Journal Citation Reports. ISI continues to explore new perspectives on the role, content and significance of the more than 21,000 journals we index, to inform researchers about the optimal venues for their papers. One of these is an indicator of national orientation.

Just as quadrupeds include both horses and turtles, the class 'scholarly and scientific journals' encompasses variances in multiple dimensions. Journals are heterogeneous in purpose, publication model, target audience, language, visibility and influence. For example, a journal may be subscription-based, open access or hybrid. It may be English-language or offer content in other languages. It may be aimed at researchers working at the frontier of a field, or delivering key findings to practitioners, such as clinicians or others who publish infrequently. In mission, audience and reach, it may serve an international, regional, national or local community.

Various approaches could be used to identify journals local to a particular region. One simple option is to look for geographic references in the name of the journal. Examples include the *Korean Journal of Applied Statistics* or the *Moscow University Mechanics Bulletin*. However, the *New England Journal of Medicine*, an internationally influential journal, demonstrates the flaw in this approach.

Similarly, publication language might offer clues – but not for journals published in English which has become the international language of research. There are also anomalous journals, such as the *International Journal of Clinical and Experimental Medicine*: published by an American publisher in English, and yet more than 95% of its content is from Chinese-affiliated researchers. This 'international journal' is therefore 'analytically' local in all but name, language and the geographical location of its publisher.

This suggests another approach for identifying local journals: the proportion of output from the lead contributing country/region. This was proposed by Moed (2005) as an indicator of national orientation (INO). Moed et al. (2020) extended the idea and proposed two indicators: INO-P for the most prolific country/region publishing in the journal; and INO-C for the most prolific country/region citing the journal.

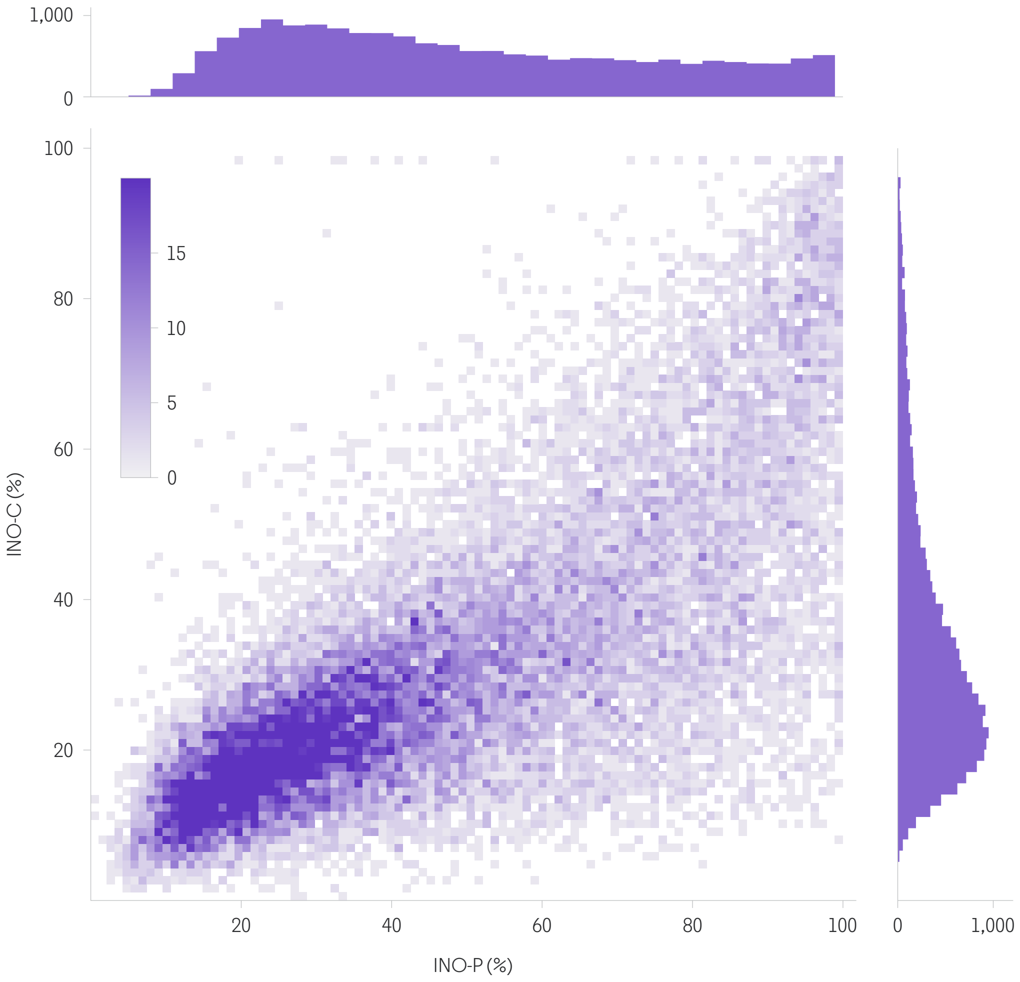
If we apply this idea to the more than 21,000 journals in the 2022 edition of the Journal Citation Reports, focusing on original academic papers (articles and reviews) and aggregating data from 2017 to 2021, we find a wide variety of values for these indicators. Some journals are highly localized to a particular country/region, with INO-P and INO-C values over 90%, while most are more international (Figure 3). Several features stand out. More than 80% of journals have a higher INO-P than INO-C. Many journals have an INO-C below 40%, suggesting citations coming from a broad mix of countries/regions, with a long tail up to 100%. And while many journals also have an INO-P below 40%, the distribution flattens out as INO-P rises above 60%.

The implication here is that while many journals may include content predominantly authored by researchers from a single country/region, many of those journals receive citations from further afield. The influence of these 'local' journals is therefore an important part of the international, indeed global, research ecosystem, helping to surface research results more widely.

No specific threshold value of INO-P or INO-C identifies a journal as being local. A spread of high thresholds could potentially be valid benchmarks for different analyses. For example, Table 2 lists the 10 countries/regions with the greatest number of local journals where INO-P and INO-C are both above 75%. Other thresholds would lead to a different result. Russia has the most local journals, which is not surprising given many Russian researchers tend to publish their results in Russian-language journals. The United States and Mainland China also have many local journals, partly because their researchers publish far more than those of any other countries/regions.

Table 2 highlights a deficiency of the two INO indicators: they do not consider journal content as a share of output from each country/region. While the U.S. is only third in the table it is also the lead publisher in 8,061 journals in the JCR: more than any other country/region, and more than the number of journals where it ranks second, or third, and so on.

Figure 3: The distribution of the percentage of papers published (INO-P) or cited (INO-C) by the most frequent author country/region for each journal indexed in Journal Citation Reports data from 2017-2021.



The distribution of authorial and citing countries/regions for the *New England Journal of Medicine* and the *Korean Journal of Applied Statistics* (Table 3) provide a clearer overview of the publishing profile of these two journals than INO-P or INO-C on their own, listing the 10 countries/regions contributing most frequently to these two journals between 2017 and 2021, with the percentage contribution for authored papers and citations to papers.

As shown in Figure 3, most journals indexed in the Web of Science have at most around 20-40% of authors (INO-P) and around 20-40% of citations (INO-C) attributed to just one country/region. In other words, most journals have international content. Some journals have a much more local authorial spread (high values of INO-P) but the geographical distribution of citations is often global, even for these journals. In other words, even local journals have international value. The next step in analysis may be best directed to regional and cultural groupings, to determine if there is an intermediate type between local and global.

Table 2: The number of local journals in the 2022 Journal Citation Reports by country/region based on INO-P and INO-C exceeding 75%.

Country/region	Number of journals
Russia	190
Brazil	171
U.S.	128
Mainland China, Hong Kong and Macau	73
Spain	62
Turkey	61
Germany	60
Ukraine	29
Australia	28
Poland	27

Table 3: The 10 countries/regions contributing most frequently as either authors or sources of citations either to *The New England Journal of Medicine* or to *The Korean Journal of Applied Statistics* (2017-2021). Percentages can total more than 100% due to country/region collaboration on papers.

The New England Journal of Medicine			The Korean Journal of Applied Statistics		
Country/region	Papers (%)		Country/region	Papers (%)	
	Published	Citing		Published	Citing
U.S.	82.8	40.1	South Korea	100.0	82.7
U.K.	27.6	9.9	U.S.	1.8	8.0
Canada	19.7	6.4	Canada	0.3	0.2
Germany	18.6	8.1	Azerbaijan	0.3	0.0
France	18.0	5.7	Singapore	0.3	0.0
Australia	14.5	4.7	Mainland China, Hong Kong & Macau	0.0	5.7
Italy	14.1	7.8	India	0.0	5.1
Spain	12.2	4.3	U.K.	0.0	2.1
Netherlands	11.7	4.3	Malaysia	0.0	0.8
Switzerland	9.4	3.2	Pakistan	0.0	0.6

04 Collaboration-CNCI
– Ross Potter

In *Profiles, not metrics* we showed how Impact Profiles revealed the spread of citation impact in a set of publications, allowing the proper comparison of numbers of high- and low-cited papers which would be masked by a single 'average' metric. Another hidden component is the influence of well-cited internationally co-authored papers.

Most academic research in the 1980s was conducted in and published by a single country/region. The U.K., for example, had an international co-author on fewer than 10% of its papers. Since then, research has become internationally collaborative (e.g., Narin et al., 1991; Adams, 2013), a trend that continues to grow (Adams et al., 2019). Collaboration is generally seen as positive and sometimes as a clear necessity (e.g., COVID-19, particle physics). However, for scientometric analysis, it can obscure the contributions of individual countries/regions, institutions and researchers.

Highly multilateral collaborations tend to be more highly cited (e.g., Narin et al., 1991; Glänzel and Schubert, 2004; Adams et al., 2019). Consequently, the average CNCI for any entity may become 'skewed' by highly multilateral papers with exceptional and erratic citation counts (Adams et al., 2022). Analysis that is blind to this effect could disproportionately influence research policy and decision making, particularly if metrics users do not understand the data context (Szomszor et al., 2021).

Methods of assigning credit across collaborative authors include full counting, fractional counting, first-author weighting and others summarized by Gauffriau (2021). No strong preference has emerged in favor of any method, possibly because continuing problems are evident for all. Weighting is value laden (First author significance? All authors equal? Equitability across disciplines?) and the 'collaboration effect' is hidden under the mask of the newly derived metric.

To address this conundrum, ISI formulated a new variant of the traditional CNCI: Collaboration, or 'Collab'-CNCI (Potter et al., 2020, 2022). The core reasoning is that, if collaboration is now integral to research and a key driver of innovation, then it should be incorporated directly and reported transparently in analyses.

Collab-CNCI follows the traditional CNCI approach but, crucially, an additional type is used for normalization. This considers collaboration complexity. Model analyses suggest that five collaboration types provide a balance between two desirable attributes: deconstruction of differences; and simplicity of interpretation. The five types are: domestic (single institutional); domestic (multi-institutional); international bilateral; international trilateral; and international quadrilateral-plus collaboration. Articles with authors from four or more countries/regions comprised about 2% of all articles in the Web of Science between 2009 and 2018.

We can compare CNCI values for the same set of papers calculated via three methods:

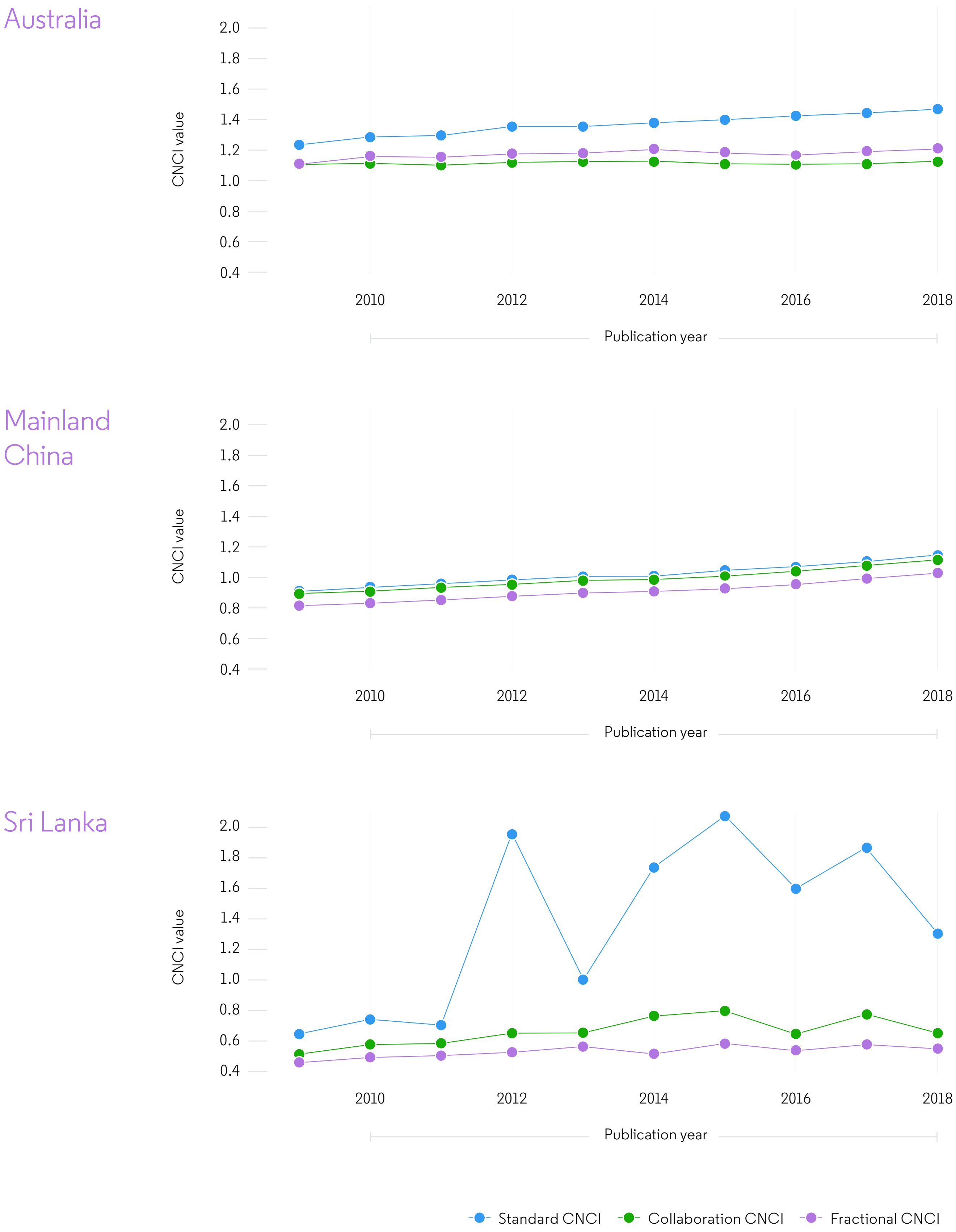
(a) standard category-based normalization;

(b) a fractional method based on Waltman and van Eck (2015) where an entity's share on a research paper is made equal to the fraction of the author addresses from that entity; and

(c) the new ISI collaboration approach.

Analyzing three different research economies (Figure 4) we find that fractional and collaboration methods produce similar results despite the very different methodology, but for Australia both values are lower than that of standard CNCI due to fractionating collaborative papers. For Mainland China, the standard and both collaboration approaches are almost identical, suggesting that its CNCI is not collaboration dependent. For Sri Lanka, however, the standard approach delivers far higher values, is more variable than other methods and is volatile. Normalizing by collaboration type smooths this variation.

Figure 4: A comparison of standard, collaboration and fractional CNCI for three countries/regions over a period of 10 years.



Australia

Mainland China

Sri Lanka

Collaboration is generally seen as positive and sometimes as a clear necessity (e.g., COVID-19, particle physics). However, for scientometric analysis, it can obscure the contributions of individual countries/regions, institutions and researchers.

This comparison tells us that while Sri Lanka appears to outperform both Australia and Mainland China over several years when using a standard CNCI methodology, this is no longer the analytical outcome when normalizing by collaboration type or author credit. Sri Lanka's standard CNCI values are, therefore, likely driven by what is normally a concealed factor for which an informed result requires a different analytical approach.

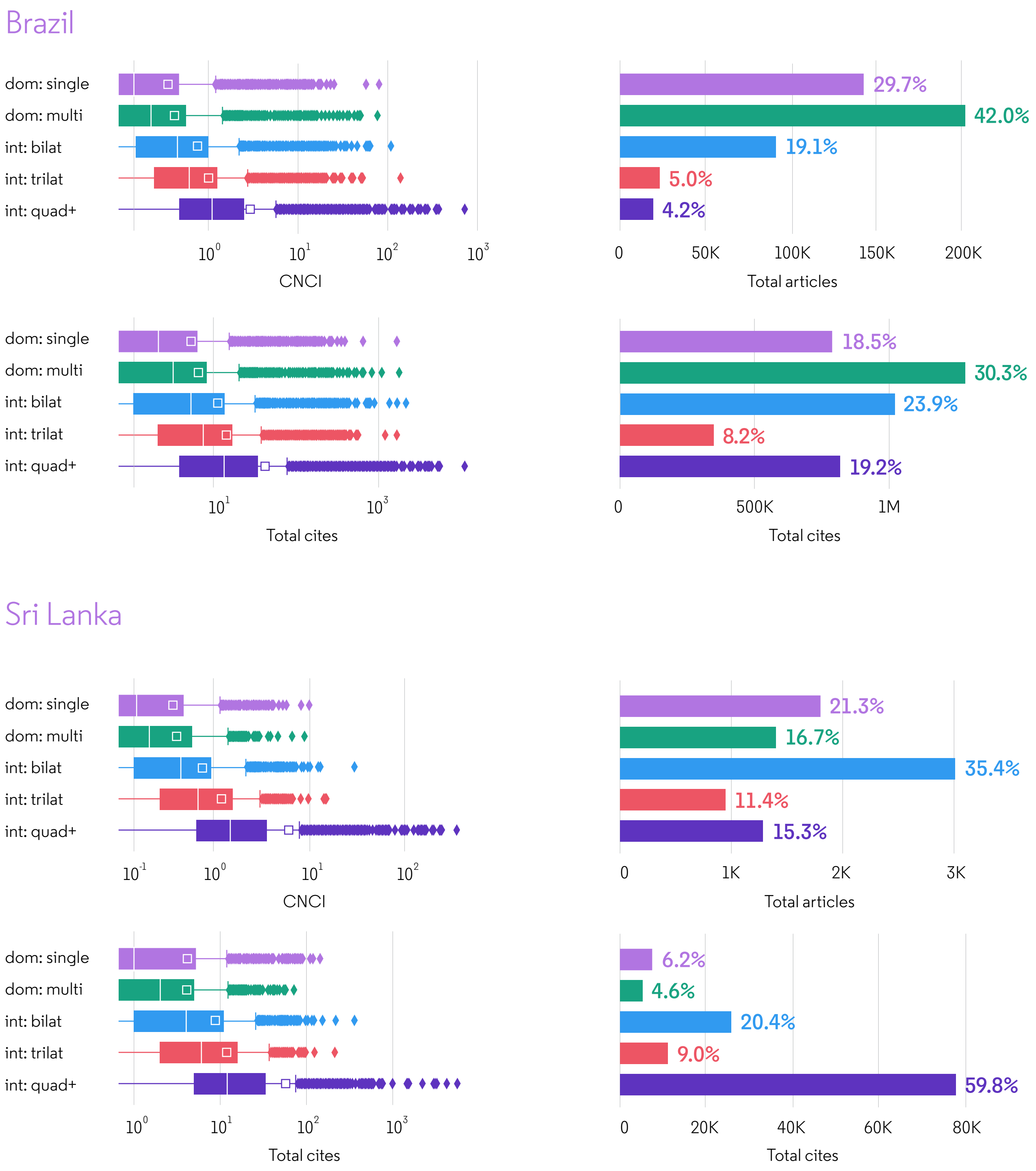
This difference in approach comes from the deconstruction enabled by Collab-CNCI. A potential criticism of Collab-CNCI as a summary indicator is that, like the standard and fractional methods, it only offers a single value snapshot of an entity's performance. However, the use of collaboration types allows a detailed deconstruction of an entity's research output for in-depth comparisons between collaboration types as well as between peer institutions or countries/regions. Data for Brazil and Sri Lanka (Figure 5) indicate that total citations and, consequently, CNCI values tend to increase as research becomes more (internationally) collaborative (left side of figure: boxplots). However, the data also show a disparity between article and citation share for each group (right side of figure: bar plots).

Brazil's research is overwhelmingly domestic (~73% of articles), likely due to its large research economy, but this research output accrues less than 50% of all its citations. Conversely, Sri Lanka's research is mainly internationally collaborative (~62% of articles), likely due to its small research economy, and that collaboration is responsible for ~90% of its citations.

We can expand the Collab-CNCI approach further by plotting values for the five collaboration types for individual institutions and over periods of time. Our publication (Potter et al., 2022) provides examples where Collab-CNCI shows domestic output performing (relative to its peers) better than internationally collaborative work.

Collab-CNCI retains the clear platform built on well-established, widely understood, standard CNCI methods. Additionally, comparisons show that the standard and fractional approaches can be used to complement this collaboration approach (Potter and Kovač, 2023), further strengthening analysis and understanding. The interpretation of collaboration type is a further argument in favor of profiling publication portfolios rather than relying on a single summary metric. Impact Profiles were featured in *Profiles, not metrics* and can be applied at several levels including institution and country/region, or even research funders. They are now part of the InCites™ product, supporting greater understanding of research activity and thereby better informing research management and policy decisions.

Figure 5: Article and citation data (2009-2018) for Brazil and Sri Lanka deconstructed by collaboration type.



Note different scales: The box plots show a spread of one standard deviation about a mean value and the individual outliers. Vertical lines within the boxes are median values; white squares, not necessarily within the boxes, represent the mean.

🟪 Domestic single institution 🟩 Domestic multi-institutional 🟦 International bilateral

🟥 International trilateral 🟪 International quadrilateral plus

05 Research Fronts
– Jonathan Adams

A limitation to the information we acquire through bibliometrics is that analysis inevitably looks back in time: via citations to prior papers about earlier projects based on original ideas. However, resource management and policy decisions cannot rely on data about past achievements. Better support would come from looking forward or close to the edge of research: the Research Front.

The history and development of Research Front technology is described in detail in our 2020 Global Research Report, *Identifying Research Fronts in the Web of Science* (Szomszor et al., 2020). Henry Small, former Director of ISI, showed that Research Fronts could be identified through recent citations to papers that were themselves among the 1% most frequently cited in their subject category (Small, 2006; Pendlebury, 2013). Highly cited papers have had exceptional influence, or 'impact', and are associated with researchers like Nobel Prize winners. However, instead of looking at those papers, Small asked questions about the citing papers, which inevitably must be more recent (Small, 1973). In particular, if a recent paper cites several highly cited papers also cited by another recent paper, then those two new papers must be focusing on the same 'next step'. A set of such papers looks like an intriguing new development: a 'Research Front', the identity of which can be established by examining the cited (core) and citing documents (Glänzel and Thijs, 2012).

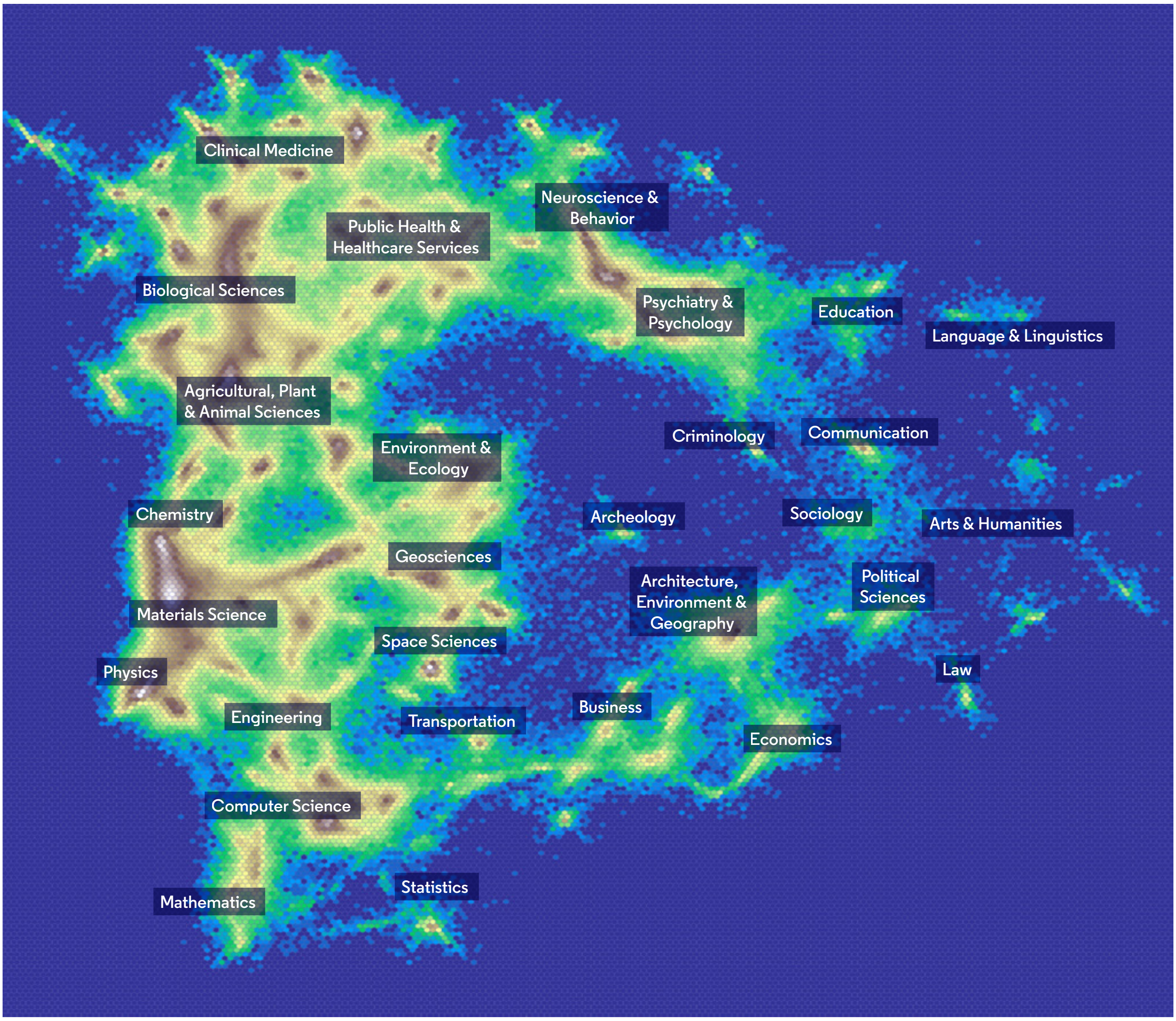
The underlying concept has been thoroughly tested and proven by both the Chinese Academy of Sciences and the Japanese Science and Technology Agency which analyze all the Web of Science data to underpin their research policy and planning. Now, we need a tool that can produce not just results but their visual presentation. Here, we describe approaches under consideration by ISI.

Approach 1: Global map

A first step in enabling interpretation of research activity is to locate the work in the wider research landscape. In the Web of Science this is facilitated by well-established and journal-based subject categories. Publications are usually compared with similar papers from a journal set covering a recognizable area of research and, often, extensively cross-referencing one another. Research Fronts are about innovation, often sparked by cross-disciplinary connections, so conventional categories are a less authoritative guide. We need a more complete map as a starting guide.

There are many approaches to creating a global map of science (Börner, 2010). ISI focuses on the recent (last two years) papers that co-cite the same highly cited papers (1% most cited) of the previous four years. Where a cluster exceeds a (arbitrary) threshold, we have a Research Front grouped by a common dependency. We then create a heat-map of current Research Fronts, shaped by cross-citations. This looks like an archipelago with mountain peaks that represent strong cores around major disciplines, with lowlands on disciplinary margins, some channels separating distant areas but rarely any isolated atolls. The map is conceptually familiar for most researchers and enables rapid location of any other linked information (Figure 6).

Figure 6: A visual representation of a global set of Research Fronts linked by common references to highly cited papers. Data are taken from papers indexed in the Web of Science for 2014-2019. Peaks of concentrated activity can be identified as the core of major disciplines, enabling rapid orientation within the 'landscape'.



We can use metadata associated with each research paper to ask about the identity of authors and institutions in the clusters.

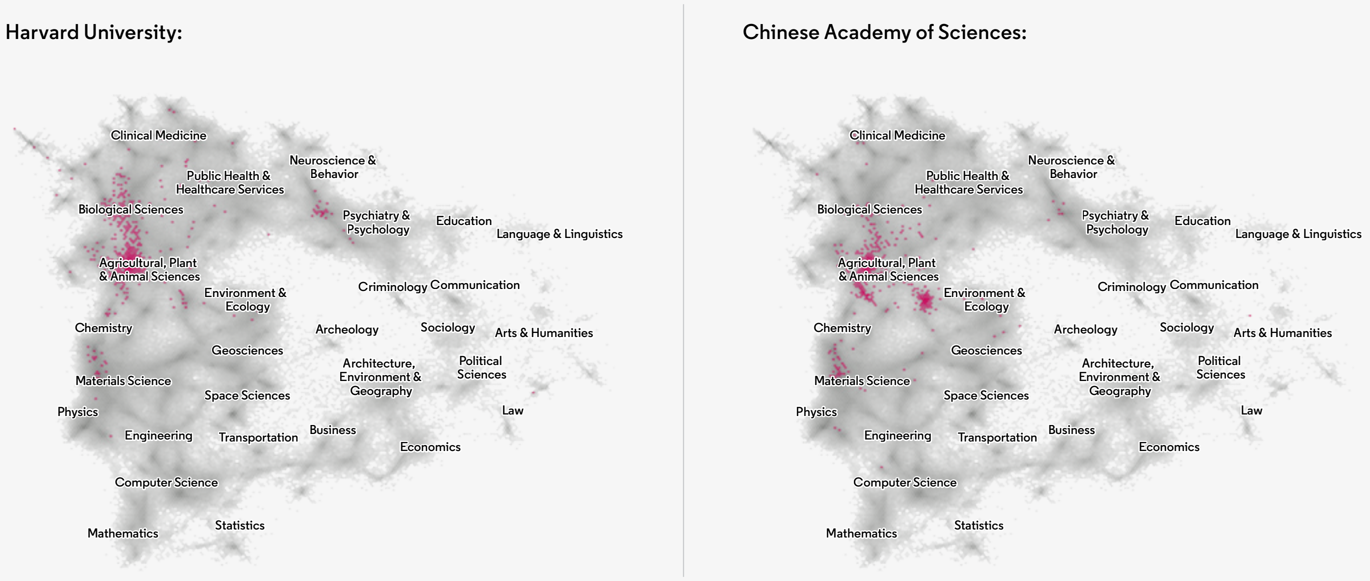
When we drop the papers from a single front onto our map, do they form a coherent cluster, or perhaps several small clusters, linked by a common theme? We find that the fronts immediately make good sense and some indeed do link sub-clusters near more than one disciplinary peak.

We can use metadata associated with each research paper to ask about the identity of authors and institutions in the clusters. We can also 'reverse the horses' and ask where the papers for any institution are located, identifying the groups contributing to innovative areas. We can also see how many institutional papers form the core, highly cited papers linked to each front.

For example, papers about CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), the core mechanism in a bacterial defense system adapted as the basis for CRISPR-Cas9 genome editing technology, are clustered around the biomedical peaks in the north-west of the archipelago. By contrast, papers about research on energy transitions spread between Engineering, Geography and Environment (see Figure 6): unsurprisingly, a highly cross-disciplinary area.

The next step reveals differences in institutional coverage. For example, Harvard University and the Chinese Academy of Sciences, two leading research institutions involved in CRISPR research, focus on overlapping but distinct parts of the research field (Figure 7).

Figure 7: The location of CRISPR papers by authors from specific institutions, illustrating the contrast in their research focus.



Harvard University:

Chinese Academy of Sciences:

Approach 2: Topic-focused clustering

A separate approach starts from a research management focus on specific research. The interest might be in evaluating, managing and investing in research capacity in a key institutional research strength or a potential research program. We are then less interested in the broad map of science and want instead to look in detail at the target area, so we start from a selective data.

The example we selected in our September 2020 report on Research Fronts was that of Geosciences. As before, we analyzed the citation links between the highly cited core papers and the papers that cited two or more of them, using the Geosciences category of Essential Science Indicators and a single target year (2017) for which there were 53,040 relevant articles associated with Research Fronts. We mapped similarity as measured by bibliographic coupling and constrained the visualization for convenience to fit a regular disc. Community detection software identified clusters of similar articles that can be highlighted in the map and reviewed to apply labels.

With a map of this kind, we can apply filters to see where entities (institutions, countries, funders acknowledged in papers) appear on the topic map. The combinaton of entity and subject maps reveals the source of core papers and how these relate to topics. Some groups of core papers are themselves clustered whereas other topic areas have few highly cited papers. Long connections across the topical landscape also emerge, connecting knowledge from different topics which in itself may signal an important innovation. We plan to implement this in our products in the near future.

The visualization of Research Fronts shows us hotspots of current research interest: innovations linked to core highly cited papers. This information can throw light on our research portfolios, our research strategy and our investment planning. It can identify exciting new areas and directions in which young researchers might head. It looks forward rather than reflecting on past achievements.

The visualization of Research Fronts shows us hotspots of current research interest: innovations linked to core highly cited papers.

06 Conclusions

Visualization of the publication and citation data 'unpacks' the information that is lost when an 'average' or other single-point metric is used as a summary. It can reveal new information, point to important additional questions, demonstrate the reasons for an unexpected outcome, and assist in next steps for planning and investment by improving user knowledge.

The detection of excessive self-citation is an important early step in a program of addressing the challenge of research integrity. The published ranges of self-citation also serve as a guide for research managers who review these issues. These indicators are already available to Web of Science users.

Diversifying our view of journals by looking at their qualitative as well as quantitative characteristics is an important reminder of the original purpose of reporting research to those who can make best use of innovative knowledge. We are investigating how relevant markers of local and international significance could feature in our products.

The deconstruction of citation impact by taking collaboration into account provides better analyses of research performance and a more informed interpretation of how the indicators actually report that performance. Collab-CNCI has been widely accepted and we are exploring how it will be featured in future product development.

ISI needs advice from potential users on how best to deliver this information. What presentation, indicators, tables and graphics will best address the questions that researchers and managers want to ask, and what will deliver the information that enables them to act?

The future integration of Research Front analysis into Clarivate products is an important goal for us and we are seeking community input before development begins – from research organizations and policy units as well as individual researchers. We want to discuss the presentation that would be most effective in addressing their questions and delivering the information that would enrich research management information.

References

Adams, J., McVeigh, M., Pendlebury, D. and Szomszor, M. (2019) Profiles, not metrics. Clarivate Analytics, London, UK.

Adams, J. (2013). The fourth age of research. Nature, 497, 557–560. DOI: 10.1038/497557a

Adams, J., Pendlebury, D. A., Potter, R. and Szomszor, M. (2019). Multi-authorship and research analytics, Clarivate Analytics, London, UK, ISBN 978-1-9160868-6-9.

Adams, J., Pendlebury, D. A. and Potter, R. (2022). Making it count: Research credit management in a collaborative world, Clarivate, London UK. ISBN 978-1-8382799-7-4

Aksnes, D.W., Langfeldt, L. and Wouters, P (2019). Citations, citation indicators, and research quality: an overview of basic concepts and theories SAGE Open, 9(1), 1-17. DOI: 10.1177/2158244019829575

Börner, K. (2010). Atlas of Science: visualizing what we know. MITB Press, Boston USA. Pp 1-288. ISBN↱978-0262014458

Garfield, E. (1955). Citation indexes for science. Science, 122(3159), 108–111. DOI: 10.1126/science.122.3159.108

Gauffriau, M. (2021). Counting methods introduced into the bibliometric research literature 1970–2018: A review. Quantitative Science Studies, 2(3), 932–975. DOI: 10.1162/qss_a_00141

Glänzel, W. and Schubert, A. (2004). Analyzing scientific networks through co-authorship. In Moed, H. F., Glänzel, W. and Schmoch, U. (eds.), Handbook of Quantitative Science and Technology Research: the use of publication and patent statistics in studies of S&T systems, Dordrecht: Kluwer Academic Publishers, pp 257-276.

Glänzel, W. and Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. Scientometrics, 91, 399-416. DOI: 10.1007/s11192-011-0591-7

Ioannidis, J.P.A., Boyack, K. and Wouters, P. Citation Metrics: A Primer on How (Not) to Normalize," PLoS Biology, 14(9): 1-7, e1002542. DOI: 10.1371/journal.pbio.1002542

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 2005–2019. PLoS ONE, 5(4), e0231735.

Moed. H.F. (2005). Citation analysis in research evaluation, pp. 1-348. Springer, Dordrecht. DOI: 10.1007/1-4020-3714-7

Moed, H.F. de Moya-Anegon, F., Guerrero-Bote, V. and C. Lopez-Illescas. (2020). Are nationally oriented journals indexed in Scopus becoming more international? The effect of publication language and access modality. Journal of Informetrics, 14(2): 101011. DOI: 10.1016/j.joi.2020.101011

Narin, F., Stevens, K. and Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. Scientometrics, 21, 313–323.

Pendlebury, D.A. (2013). Research fronts: In search of the structure of science. In, Research Fronts 2013: 100 Top-Ranked Specialties in the Sciences and Social Sciences, C. King and D.A. Pendlebury (eds.), Thomson Reuters, pp 26-31.

Potter, R. W. K. and Kovač, M. G. (2023). Tracking Category Normalized Citation Impact (CNCI) changes: Benefits of combining standard, collaboration and fractional CNCI for performance evaluation and understanding. 19th International Conference on Scientometrics and Informetrics, abstract #126.

Potter, R. W. K., Szomszor, M. and Adams, J. (2020). Interpreting CNCIs on a country-scale: The effect of domestic and international collaboration type. Journal of Informetrics, 14(4), 10175. DOI: 10.1016/j.joi.2020.101075.

Potter, R. W. K., Szomszor, M. and Adams, J. (2022) Comparing standard, collaboration and fractional CNCI at the institutional level: Consequences for performance evaluation. Scientometrics, 127, 7435-7448. DOI: 10.1007/s11192-022-04303-y.

Small, H. (1973). Co-Citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265-269. DOI: 10.1002/asi.4630240406

Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68, 595–610. DOI: 10.1007/s11192-006-0132-y

Szomszor, M., Pendlebury, D.A. and Adams, J. (2020). How much is too much? The difference between research influence and self-citation excess. Scientometrics, 123 (2), 1119–1147. www.link.springer.com/article/10.1007/s11192-020-03417-5

Szomszor, M., Pendlebury, D. and Rogers, G. (2020) Identifying Research Fronts in the Web of Science: From metrics to meaning, Clarivate, London UK. ISBN 978-1-9160868-8-3

Szomszor, M., Adams, J., Fry, R., Gebert, C., Pendlebury, D. A., Potter, R. W. K. and Rogers, G. (2021) Interpreting bibliometric data. Frontiers in Research Metrics and Analytics, 5, 30pp. DOI: 10.3389/frma.2020.628703.

Waltman, L., and van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. Journal of Informetrics, 9, 872–894. DOI: 10.1016/j.joi.2015.08.001

About the Global Research Report series from the Institute for Scientific Information

Our Global Research Reports draw on our unique industry insights to offer insights, analysis, ideas and commentary to enlighten and stimulate debate.

Each one in the series demonstrates the huge potential of research data to inform management issues in research assessment and research policy and to accelerate development of the global research base.

Advice on the use of the standard methodology and information about comparative institutional analyses used in this report is available.

Email: **isi@clarivate.com**

Previous reports include:

Profiles, not metrics

Research integrity: Understanding our shared responsibility for a sustainable scholarly ecosystem

Research assessment: Origins, evolution, outcomes

Identifying Research Fronts in the Web of Science: From metrics to meaning

Making it count: Research credit management in a collaborative world

Download here:
www.clarivate.com/isi

About Clarivate

Clarivate™ is a leading global information services provider. We connect people and organizations to intelligence they can trust to transform their perspective, their work and our world. Our subscription and technology-based solutions are coupled with deep domain expertise and cover the areas of Academia & Government, Life Sciences & Healthcare and Intellectual Property. For more information, please visit **clarivate.com**.

The Web of Science™ is the world's largest publisher-neutral citation index and research intelligence platform. It organizes the world's research information to enable academia, corporations, publishers and governments to accelerate the pace of research.

Need to evaluate research at your organization?
Contact us to find out how Clarivate can help:

clarivate.com/contact-us