Under-reporting research relevant to local needs in the South

Database biases in rice research

Ismael Ràfols[(1)(2)], Tommaso Ciarli[(2)], Diego Chavarro[(2)]

(1) Ingenio (CSIC-UPV), Universitat Politècnica de València, Spain
(2) Science Policy Research Unit (SPRU)
University of Sussex, United Kingdom

1 Context

Science is increasingly expected to help address grand challenges and societal problems such as tackling obesity, climate change, and pandemics. It is important to understand the different research options than can help to tackle these problems, and to consider the directions in which scientific research should be developed. Bibliometrics can provide helpful tools for developing representations of the existing "supply" of science. However, these representations depend on the data and methods used. Bibliometric tools or indicators, therefore, often reflect choices made in data collection and treatment. Conventional bibliometric analyses are biased against non-English languages, applied research, the social sciences and humanities, and interdisciplinary research. This study investigated the biases of available databases in the representation of research topics, particularly those related to developing countries and topics potentially relevant to disenfranchised populations.

In spite of notable differences between the Web of Science (WoS) and Scopus, comparisons between these two main databases have produced similar rankings of publication production by country in different fields. When intergovernmental agencies benchmark science, these two databases continue to be used. However, the partial coverage of these main commercial databases may lead to serious misrepresentation of science in developing countries. There is a need to improve scientometric indicators to properly evaluate global science.

Agricultural research is an important endeavour in developing countries. Scientists are under considerable pressure from stakeholders to solve local problems rather than contribute to the development of "universal" knowledge. This study focused on rice because it is a staple crop for millions of people across the world. Rice research was at the core of the green revolution and it sparked numerous controversies relating to the impoverishment of diets, overuse of water, exhaustion of soils, and pollution. Given the applied orientation of rice research, the local specificity of the topics, and the relative lack of relevance of the topic for many developed countries, rice research is an interesting case to test the extent of coverage by the main bibliographic databases.

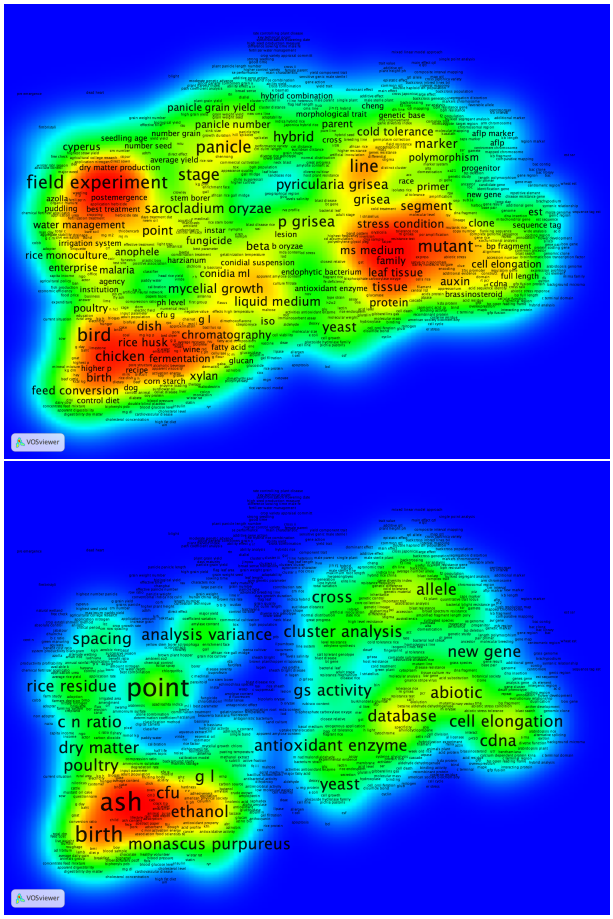
This study investigated biases in rice research by comparing coverage of WoS and Scopus with CAB Abstracts (Centre for Agriculture and Biosciences International, CABI), which is a specialized agriculture and environment database with a broader coverage of developing countries. The second contribution was to describe a substantive bias in coverage of different research topics. Third, the potential effects of these biases on policy were explored.

2 Empirical Approach and Main Findings

WoS and Scopus are well-known databases provided by large information and publishing companies, Thomson Reuters and Elsevier, respectively. CAB Abstracts is a database focused on environment and agriculture. It is run by CABI, an intergovernmental, not-for-profit organization with a mission to provide information to practitioners and apply scientific expertise to solve problems in agriculture and the environment.

Publications on rice were manually downloaded from WoS by searching "rice" or "oryza" in the topic field. Scopus records were also manually downloaded by searching in the title, abstract, and keywords fields ("rice" OR "oryza"). Similarly, documents with "rice" or "oryza" were searched in the title and abstract fields of CAB Abstracts. The records of the different databases were matched with multiple matching algorithms. The analysis was carried out using the program VantagePoint, the statistical package R, and the visualization program VOSviewer.

First, descriptive statistics were used to provide information on the number of publications by document type, language, and year. Second, the analysis compared the number of publications by country and the research topic of the publications. The author affiliation was used to retrieve information on the country. An important caveat is that CABI only reports the affiliation of the first author. In the case of WoS and Scopus, the affiliations of all authors are included. As a result, countries numbers will tend to be higher in WoS and Scopus. This effect was not corrected. The error was estimated (using small document samples) as 10–30% over-representation, depending on country. Third, rice research was clustered into six fields on the basis of the co-occurrence of terms in abstracts and titles. Each publication was then fractionally assigned to a cluster and estimates were made of the number of publications per cluster and country.



Document characteristics : year, type, and language — There were differences between the documents retrieved from each database. Trends over time showed that CABI historically had a much broader coverage than WoS and Scopus. Before the 1980s, coverage by WoS and Scopus of publications on rice was very limited. CABI showed a great increase in rice publications from the post-war until the mid-1970s, particularly after the mid-1960s driven by the green revolution. Post-war expansion was followed by a period of slow growth from 1975 until 2000, when renewed growth was observed (perhaps in coincidence with the advent of genomic studies). Since the mid-1990s, WoS and Scopus have been catching up with CABI, and by 2012 WoS had reached 80% of CABI and Scopus 86% in total number of publications, although with substantial non-overlapping coverage. We carried out the rest of the study only for the period 2003–2012, which was more relevant for policy analysis.

Each database classifies documents into different type categories. We downloaded all document types. Of these, journal articles were the dominant type (94% in CABI, 95% in Scopus, and 91% in WoS). The second most important document type was

conference proceedings and papers (6.2% CABI, 4.6% in Scopus, and 7.0% in WoS). In terms of language, CABI was much more comprehensive, with 25% of publications in languages other than English, compared with 11% of Scopus and 4% of WoS.

Comparison of coverage across countries — Researchers in India, China, Japan, and the United States (US) publish the most rice research. China's publications on rice have sharply increased in the last 20 years, while the shares by India, Japan, and the US have decreased. All three databases agree on these trends. However, there were major differences in the overall proportion of publication assigned to each country in each database. In the case of CABI, India was the most productive country until it was caught by China in 2004. Whereas India's publications made up 18% of the total in 2003–2012 according to CABI, they represented 11.2% and 9.6% in Scopus and WoS. Similarly, China's publications were 26% of CABI's publications, but only 22% and 18% in Scopus and WoS. In comparison, US publications were only 6% in CABI, but 10% and 11% in Scopus and WoS. The differences in coverage between databases have narrowed in recent years, but there is still almost a two-fold difference in the share of publications assigned by CABI and WoS for the US and India.

The percentages of publications from countries with a smaller number of publications on rice showed that developing countries had much higher coverage in CABI. Industrialized countries had a much higher coverage in WoS. Middle-income countries in Asia scored similar shares in all three databases. From the analysis, it follows that WoS and Scopus cover research published in North America and Europe ; whereas, CABI is much more comprehensive in developing countries. As a result of CABI's larger coverage, the relative contribution of Western countries to scientific production on rice is much smaller than is usually acknowledged when using standard publication databases such as WoS and Scopus.

Comparison of coverage across research topics — WoS has a wider coverage of the biomedical sciences ; whereas, Scopus has relatively better coverage of the social sciences and humanities. CABI is expected to have a more comprehensive coverage of agronomy and environment. The problem for making a comparison across disciplines or topics is that a shared disciplinary classification across the three databases is needed, but each database provides its own classification. To have a shared topic view across the three databases with a classification relevant to rice research, all publications with abstract for the period 2002–2012 (78,225 articles) were pooled, and terms were clustered using the freeware VOSviewer. Given potential problems with this bottom-up method of classification, an alternative clustering algorithm was also used. The clusters differed slightly (e.g., the consumption cluster split between food processing and human nutrition), but the overall findings were consistent.

Six research topics were found : a consumption cluster that includes human nutrition (e.g., diet) and food processing (e.g., starch) ; a cluster of productivity and plant nutrition that includes research to increase crop yield, and socioeconomic issues ; a molecular biology cluster that includes genomics and transgenic research ; a genetics cluster that appears to reflect hybridization approaches ; a cluster describing plant characteristics (e.g., panicle and grain length) ; and a cluster related to diseases, pests, and related efforts to protect the rice plant.

A closer look at the coverage of the different databases per research topic showed that CABI's coverage remained above 80%, except for the topic of consumption (65%). The coverage of Scopus was above 70% for molecular biology and consumption, but in the range of 30–50% for rice production, pests, and plant characteristics that are more directly related to improving yields. Similarly, WoS coverage was very low (20–30%) for production-related topics, but higher for molecular biology and consumption (50–60%).

Why is there such a disparate degree of coverage between topics by mainstream databases? An initial hypothesis is that the topics with low coverage are those where developing countries publish relatively more, and those with high coverage are those where developed countries publish more. Low-income countries tend to focus on productivity; whereas, most developed countries focus on molecular biology. Interestingly, some middle-income countries (e.g., Thailand and Malaysia) had a high percentage of publications in consumption, which had high coverage by WoS and Scopus. In general, there is a loose relation between lack of coverage by Scopus and WoS, and most of the topics relevant to developing countries, with the exception of consumption.

Whatever the reason, there is no doubt that the low coverage of many low- and middle-income countries in the mainstream databases results in a very distorted perspective of the research they carry out. For example, the publication portfolio of rice for Iran in CABI shows expertise in rice production and consumption, with some work on genetics, but in WoS, only the consumption side shows up. Similarly, the publication portfolio of rice for Cuba is atypical, but interesting, because it shows that Cuban publications in molecular biology are not covered in WoS or Scopus.

There are a number of limitations in the empirical strategy adopted. At this stage, the data have not been corrected for the fact that CABI only reports the affiliation of the first author. Second, although CABI coverage of publications is possibly the largest on a subject such as rice, publications still represent a subset of the research actually carried out. In research on agriculture, many research outputs are not accounted for in publications and a lot of research is done by private companies and public organizations. Third, a bottom-up classification in terms of research topics is inevitably open to debate, although the same findings were obtained with a different clustering method.

3 Main messages for policy and practice

This study shows that previous assumptions about the stability of indicators across databases of scientific production may be incorrect. This case study on rice research confirms the view that publication statistics per country are very dependent on the database used. A new finding of this study on rice is that there is also a serious bias in coverage of research topic. These differences are likely to apply as well to other fields of agricultural research, particularly those related to crops less relevant in the temperate climates of industrialized nations. These results are potentially important for international organizations that work on agriculture and human development. Nevertheless, the findings do not come as a surprise given the proliferation in the last two decades of journal indexing systems at the regional level (e.g., Scielo or Redalyc)

that aim to provide visibility to local journals, often in languages other than English. These journals indexing systems exist to compensate for the fact that local science and its journals are not perceived as participating in international science.

Knowledge representation can play a significant role in framing evaluation, research strategies, and technological development policies, and this study demonstrates that topic bias in the dominant bibliometric databases deserves further conceptual and empirical discussion. This is an important issue to discuss, because biases in the representation of knowledge may inform research strategies. In particular, bibliometric reports of developing countries using dominant databases could lead to inappropriate assumptions and choices regarding domestic science bases and capabilities.

At the heart of this debate lies the question of how organizations or countries should manage and assess research priorities so as to better align science "supply" with societal needs or "demands." Because bibliometric studies are part of the governance of science and innovation, the way they represent knowledge is bound to have an effect on the understanding and prioritization of research. Judgements about the relative worth of research are influenced by dominant norms in science that tend to value more universal over local findings, pure over applied, and laboratory work over field work. The topic biases reported here may further reinforce these biases with the ultimate effect of making science and technology less appropriate for the local needs of farmers in the South. For these reasons, we believe that this apparently technical study is relevant for discussion of how the governance of science relates to inclusive innovation.

Related Resources

Arvanitis, R. and Chatelin, Y. 1988. National scientific strategies in tropical soil sciences. *Social Studies of Science,* 18, 113–146.

Velho, L. and Krige, J. 1984. Publication and citation practices of Brazilian agricultural scientists. *Social Studies of Science*, 14, 45–62.

Vessuri, H., Guédon, J.C., and Cetto, A.M. 2014. Excellence or quality ? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. *Current Sociology*, 62, 647–665.