CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Dissertation Thesis

2023 **Vít Macháček**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Vít Macháček

Chapters in a cross-country analysis of science

Dissertation Thesis

Prague 2023

Author: **Vít Macháček**

Supervisor: **Ing. Martin Srholec, Ph.D.**

Year of the defense: **2023**

Abstract

The dissertation consists of three independent chapters:

Chapter 1 – *Predatory publishing in Scopus: Evidence on cross-country differences* – presents evidence on how predatory journals have infiltrated research systems across various countries. The chapter builds on data from the Scopus database in combination with the content of the so-called Beall's list. Bas on this data we identify 324 potentially predatory journals. Then we compare the number of papers in these journals with the total number of papers for each country. The results reveal considerable heterogeneity across countries. While the most affected countries – Kazakhstan and Indonesia – published 17 % of their output in the journals indexed on Beall's list, the share was below 1 % in developed countries.

The second chapter – *Researchers' institutional mobility – bibliometric evidence on academic inbreeding and internationalization* – analyzes researchers' career paths at 1,130 universities included in the Leiden Ranking 2020. Researchers affiliated with one of these universities and publishing in 2018 were split into categories according to the affiliation stated on their earliest publication: i) *insiders* published their first papers while affiliated with the same university, ii) *domestic outsiders* started at a different university in the same country, and iii) *foreign outsiders* started publishing in different country. The share of insiders is a proxy for a university's tendency to engage in academic inbreeding – a commonly criticized practice of universities hiring their graduates. We identify a robust spatial structure. Universities in Anglo-Saxon countries and North Western Europe tend to be highly mobile, but the rest of the world often hires their own graduates. To our knowledge, this is the first scientometric evidence of academic mobility taking an institutional perspective.

The third chapter – *Globalization of Scientific Communication: Evidence from authors in academic journals by country of origin* – measures the tendency to publish research in domestic journals across countries and disciplines. The methodology determines seven indicators of globalization for each journal indexed in Scopus. The individual journals' indicators are subsequently aggregated to the country and discipline level. In developing countries, the globalization of scientific communication is high and relatively homogenous. Contrary, countries of the former Soviet bloc still heavily rely on local journals. Developing countries are in between the previous groups. The role of local journals is higher in social and health sciences than in physical and life sciences.

Declaration of authorship

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on Vít Macháček

Acknowledgments

Undoubtedly, my most tremendous gratitude belongs to my supervisor – Martin Srholec. Working together daily, writing provocative IDEA studies, and slowly converting them into academic articles has been a pleasure. The last months of my doctoral studies were less enjoyable, but Martin had a heart in the right place.

Of course, huge thanks to my wife Helena. It would be nice if the first lines of my academic text she ever reads were this acknowledgment. But she allowed all that to happen. Only together (and with her mum) we made it through all the turbulence. Also, Mikuláš and Dorotka deserve their part – especially for the evenings I did not play with them.

And of course – Petr Pleticha. He suffered a lot, especially when he wanted to work, but I had some ideas that were absolutely necessary to discuss immediately.

The staff on CWTS in Leiden also left significant traces on the manuscript. To name a few: Ludo, Nees Jan, Rodrigo, Ismael and Juan Pablo. Who knows if I would ever finish this dissertation without them.

Funding

This work was supported by the Czech Academy of Sciences for the R&D&I Analytical Centre (RaDIAC) and from the Charles University Grant Agency [no 1062119].

List of tables and figures

List of tables

List of figures

Table of contents

Introduction

“When two do the same thing, it is not the same thing.” (Czech aphorism)

Like any other human endeavor, doing science is embedded into local environments. Researchers worldwide are affected by various cultural, economic, institutional, and policy incentives (Franzoni et al., 2011). The science they produce will be affected by the reward system researchers face and more implicit factors such as cultural norms and networking opportunities (Good et al., 2015).

The dissertation’s leitmotif is the demonstration of how varying incentives shape scientific practice. All chapters reveal significant differences between countries and even whole regions. Publications in predatory journals are much more prevalent in the Global South than in developed countries. Publishing science in domestic journals is the norm in Russia, but it is rare in Denmark. Post-graduates from Anglo-Saxon universities are more likely to move to another university than post-graduates from the rest of the world.

These findings might seem disjointed. On the one hand, this is not surprising as individual chapters are based on previous IDEA policy studies (Macháček and Srholec, 2017a, 2017b, 2018, 2019, 2021), each of which answers a different research question. On the other hand, a clear motivation informs all the research questions answered in this dissertation: The desire to understand the role of public policies aimed at research.

The literature provides some understanding of the processes leading to chapters’ observed outcomes, such as how researchers decide where to submit their papers or how they make mobility decisions (Baguess et al., 2019, Horta, 2013). It also generally argues that some of the issues studied in this dissertation, such as academic mobility and internationalization, promote quality in science (Franzoni et al., 2014, Buela-Casal et al., 2006, Science Europe, 2013, Wagner and Jonkers, 2017, Royal Society, 2011, European Commission, 2018).

However, what the literature does not provide are meaningful descriptive indicators of the questions this dissertation tackles. They are either (as far as we know) not studied at large-scale and in a systematic way at all, or the evidence is over 20 years old (Zitt & Bassecoulard, 1999).

This dissertation fills these research gaps by providing large-scale, cross-country-comparable, and systematic evidence on the penetration of predatory publishers, on academic inbreeding, and on the globalization of scientific communication. This contribution can stand on its own in the literature, but it also allows fruitful future research: The scientometric data this dissertation provides can serve as a basis for country-level analyses assessing the relationship between studied phenomena and various policy mixes, but also allows to go directly to micro-level, where the supplemented data on individual journals and researchers would be suitable.

Chapter 1 – *Predatory publishing in Scopus: Evidence on cross-country differences* – studies publishing in so-called "predatory journals." The term coined by Beall (2016) refers to open-access journals charging subscription fees but not performing sufficient peer-review processes. Instead, they might publish work that would not qualify as a valid scientific paper elsewhere. Specifically, this chapter analyzes the propensity to publish in suspected predatory journals across countries.

The chapter is built on the identification of 324 journals listed by Jeffrey Beall in his blog (Beall, 2016) and indexed in the Scopus database. Between 2015 and 2017, Scopus indexed 164 thousand articles from these journals. The analysis reveals considerable heterogeneity in countries' propensity to publish in suspected predatory journals.

Large middle-income countries in Asia are most prone to predatory publishing. Among those most susceptible to predatory publishing are Kazakhstan (17 %), Indonesia (16 %), Malaysia (12 %), and India (10 %). Social and health sciences seem more vulnerable to predatory publishing than physical and life sciences.

Regression analysis identified an inversed U-shaped relationship between the propensity to publish in suspected predatory journals and GDP per capita. The peak is close to GDP per capita of India, Nigeria, and Pakistan. Also, rent from natural resources seems to be a significant determinant – suggesting that governments' unconstrained access to finance may not automatically translate into quality research.

The paper's publication triggered significant feedback from the academic community (Abramo et al, 2022, Mills and Bell 2021, Chawla, 2021). It also motivated Scopus to take action and discontinue the indexation of 97 titles identified in the paper (Scopus 2021).

Academic inbreeding is a phenomenon studied in Chapter 2 - *Studying researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization.* It is a practice of universities to hire their own graduates. According to critics, universities with high degree of academic inbreeding do not sufficiently utilize the international labor market for researchers. Consequently, inbreeding hinders the development of both universities and the researchers' careers.

The chapter proposes a novel methodology for studying the mobility patterns of universities. Researchers publishing with a university affiliation in 2018 (who also published their first papers in 2012 or earlier) are split into three categories according to where they started their career: i) *insiders* who published their first paper affiliated with the same university, ii) *domestic outsiders* who published their first paper with the university in the same country, and iii) *foreign outsiders* who started their career elsewhere. The share of insiders is used as a proxy for academic inbreeding.

The analysis covers 1,130 universities included in Leiden Ranking 2020 (van Eck et al, 2020) and uses the Dimensions database as a data source. The chapter reveals an important spatial pattern – universities in USA, Canada, and Western and Northern European countries tend to have a relatively low share of insiders – the median is between 30 and 40 %. The rest is hired on the external market. On the contrary, institutions in Eastern and Southern Europe tend to rely more on internal resources – the median share of insiders is between 70 and 80 %. '

Stratifying the sample across researchers' disciplines reveals that disciplines are not an essential determinant of the share of insiders. National and university-related factors are much stronger predictor. This finding suggests that the tendency to hire from external resources largely depends on the internal processes of universities, which reflect national incentives and cultural norms. On the contrary, disciplinary habits play only a minor role.

Studying mobility patterns has always been a challenge for scientometrists, as mobility events are hard to trace in the data. The chapter demonstrates that now the data are mature enough to deliver meaningful results.

Internationalization is also a topic for the last chapter of the dissertation – *Globalization of Scientific Communication: Evidence from authors in academic journals by country of origin.* The paper calls for analyzing globalization through the lens of scientific communication next to traditional international cooperation-based measures.

In developed countries, the "transition to the international mode of scientific communication" took place in the 1980s (Zitt et al. 1998). Publishing in international journals became the modus operandi for researchers in these countries. Our analysis confirms this. However, our evidence also shows that 30 years later, the globalization of scientific communication is still heterogeneous across countries. For example, less than 5 % of Danish social science articles are published in 25 % least globalized journals. This figure is approximately 50 % in Croatia, Slovakia, and Latvia.

Academic journals became an essential input for research evaluation. We argue that establishing a local journal ecosystem can serve as a tool for "gaming the system" and allow researchers to avoid fierce competition from abroad.

References

Abramo, G., Aguillo, I. F., Aksnes, D. W., Boyack, K., Burrell, Q. L., Campanario, J. M., ... & Waltman, L. (2022). Retraction of Predatory publishing in Scopus: Evidence on cross-country differences lacks justification. Scientometrics, 1-3.

Baguess M., Sylos-Labini M., Zinovyeva N. (2019). A Walk on the Wild Side: `Predatory' journals and information asymmetries in scientific evaluations. Research Policy. 48(2), 462-477

Bloudoff-Indelicato, M. (2015). Backlash after Frontiers journals added to list of questionable publishers. Nature, 526, 613.

Beall, J. (2016). Scholarly Open Access: Critical analysis of scholarly open-access publishing (Beall's blog) Available from https://scholarlyoa.com; shutdown January 2018, archived at https://archive.org/web/; [cited 1 April 2016].

Buela-Casal, G., Perakakis, P., Taylor, M., & Checa, P. (2006). Measuring internationality: Reflections and perspectives on academic journals. Scientometrics, 67(1), 45–65. https://doi.org/10.1007/s11192-006-0050-z

Chawla, D. S. (2021). Hundreds of 'predatory'journals indexed on leading scholarly database. Nature. https://www.nature.com/articles/d41586-021-00239-0

European Commission (2018) ERA progress report 2018. Luxembourg: Publications Office of the European Union. 13 Aug 2021. https://op.europa.eu/en/web/eu-law-and-publications/publication-detail/-/publication/5641328c-33f8-11e9-8d04-01aa75ed71a1

Franzoni, C., Scellato, G., & Stephan, P. (2011). Changing incentives to publish. Science, 333(6043), 702–703. https://doi.org/10.1126/science.1197286

Franzoni, C., Scellato, G., & Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. Economics Letters, 122(1), 89-93.

Good, B., Vermeulen, N., Tiefenthaler, B., & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. Research Evaluation, 24(2), 91–105. https://doi.org/10.1093/reseval/rvu035

Horta, H. (2013) Deepening our understanding of academic inbreeding effects on research information exchange and scientific output: new insights for academic based research. Higher Education, 65, 487-510. http://dx.doi.org/10.1007/s10734-012-9559-7

Macháček, V. and Srholec, M. (2017a). Predatory journal in Scopus. IDEA study 2 / 2017, https://idea-en.cerge-ei.cz/files/IDEA_Study_2_2017_Predatory_journals_in_Scopus/mobile/index.html#p=1.

Macháček, V. and Srholec, M. (2017b) Místní časopisy ve Scopusu, IDEA Study 17/2017. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. https://idea.cerge-ei.cz/studies/2017-17-mistni-casopisy-ve-scopusu

Macháček, V. and Srholec, M. (2018) Kde se nejvíce publikuje v predátorských a místních časopisech? Bibliometrická analýza trochu jinak, IDEA Study 4/2018. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. https://idea.cerge-ei.cz/files/PredatoriMistni/

Macháček, V. and Srholec, M. (2019) Globalization of science: Evidence from authors in academic journals by country of origin, IDEA Study 6/2019. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. http://www.globalizationofscience.com/

Macháček, V. and Srholec, M. (2020) Where do universities recruit researchers?, IDEA Study 1/2020. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. https://idea.cerge-ei.cz/files/RecruitingResearchers/

Mills, D., & Bell, K. (2021). National propensities?. Scientometrics, 1-3.

Royal Society. (2011). Knowledge, networks and nations: Global scientific collaboration in the 21st century. https://royalsociety.org/-/media/Royal_Society_Content/policy/publications/2011/4294976134.pdf

Science Europe. (2013). Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility. https://www.elsevier.com/__data/assets/pdf_file/0019/53074/Comparative-

Benchmarking-of-European-and-US-Research-Collaboration-and-Researcher-Mobility_sept2013.pdf

Scopus (2021). The importance of high-quality content: curation and reevaluation in Scopus. Available from: https://www.elsevier.com/__data/assets/pdf_file/0004/891058/ACADLIBSCARTImportance-of-high-quality-contentWEB.pdf [cited 13 June 2022]

Silver, A. (2017). Controversial website that lists ‘predatory’ publishers shuts down. Nature, 541(7638). https://www.nature.com/articles/nature.2017.21328

Van Eck, Nees Jan. (2020). CWTS Leiden Ranking 2020 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3997240

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. In Nature (Vol. 550, Issue 7674, p. 33). Nature Publishing Group. https://doi.org/10.1038/550032a

Zitt, M., Perrot, F., & Barré, R. (1998). The transition from “national” to “transnational” model and related measures of countries’ performance. Journal of the American Society for Information

Zitt, M., & Bassecoulard, E. (1999). Internationalization of communication a view on the evolution of scientific journals. Scientometrics, 46(3), 669–685. https://doi.org/10.1007/BF02459619

Chapter 1: Predatory publishing in Scopus: Evidence on cross-country differences

Co-authored with Martin Srholec.

Originally published as Macháček, V., & Srholec, M. (2021). RETRACTED ARTICLE: Predatory publishing in Scopus: evidence on cross-country differences. Scientometrics, 126(3), 1897-1921,

The original article was retracted (see Macháček and Srholec 2022a). The justification of the retraction was disputed in 2022 by 27 members of the Distinguished Reviewers Board of Scientometrics and/or recipients of the Derek de Solla Price Medal (see Abramo et al. 2022).

This is a republished version of the previous manuscript:

Macháček, V., & Srholec, M. (2022). Predatory publishing in Scopus: Evidence on cross-country differences. *Quantitative Science Studies*, *3*(3), 859-887.

Abstract

Predatory publishing represents a major challenge to scholarly communication. This paper maps the infiltration of journals suspected of predatory practices into the citation database Scopus and examines cross-country differences in the propensity of scholars to publish in such journals. Using the names of “potential, possible, or probable” predatory journals and publishers on Beall’s lists, we derived the ISSNs of 3,293 journals from Ulrichsweb and searched Scopus with them. 324 of journals that appear both in Beall’s lists and Scopus with 164 thousand articles published over 2015-2017 were identified. Analysis of data for 172 countries in 4 fields of research indicates that there is a remarkable heterogeneity. In the most affected countries, including Kazakhstan and Indonesia, around 17 % of articles were published in the suspected predatory journals, while some other countries have no articles in this category whatsoever. Countries with large research sectors at the medium level of economic development, especially in Asia and North Africa, tend to be most susceptible to predatory publishing. Policymakers and stakeholders in these and other developing countries need to pay more attention to the quality of research evaluation.

Introduction

"Predatory" (or fraudulent) scholarly journals exploit a paid open-access publication model: the publisher does not charge subscription fees, but receives money directly from the author of an article that becomes accessible for free to anyone. However, this entails a conflict of interests that has the potential to undermine the credibility of open-access scholarly publishing (Beall 2013). Authors are motivated to pay to have their work published for the sake of career progression or research evaluation, for instance (Bagues et al. 2018; Kurt 2018; Demir 2018). In return, predatory publishers turn a blind eye to any limitations of papers during peer-review in favor of generating income from authors' fees; the worst of them fake the peer-review process and print almost anything for money, without scruples (Bohannon 2013; Butler 2013).

So far, only a handful of studies have examined the geographical distribution of authors published in journals suspected of predatory practices by Beall (2016). On a sample of 47 such journals, Shen and Bjork (2015) found that the authors were highly skewed to Asia and Africa, primarily India and Nigeria. Xia et al. (2015) examined 7 pharmaceutical journals and also identified the vast majority of authors as being from Southeast Asia, predominantly India, and, to a lesser extent, Africa. Demir (2018) combed through 832 suspected predatory journals and confirmed that by far the greatest number of authors are from India, followed by Nigeria, Turkey, the United States, China and Saudi Arabia. Wallace et al. (2018) focused on 27 such journals in economics, in which the authors were most frequently from Iran, the United States, Nigeria, Malaysia and Turkey.

No matter how insightful these studies are in revealing from where contributors to suspected predatory journals originate, we still know very little about the magnitude of the problem for the respective countries and regions. India appears to be the main hotbed of predatory publishing, but in the context of India's gigantic research system, this may be much ado about little. All of the countries cited above are, unsurprisingly, quite large. Could it be that some smaller countries are actually far worse off, though they do not stand out in the absolute figures? Just how large is the propensity to predatory publishing at the national level? Which countries are most and least affected by this problem, and why?

Existing literature provides very scant evidence along these lines and the studies at hand are limited to individual countries and use different methodologies, so the results are not easily comparable. For example, Perlin (2018) found that suspected predatory journal articles accounted only for about 1.5 % of publications in Brazil, while Bagues et al. (2019) showed that around 5 % of researchers published in such journals in Italy. No study has yet examined the penetration of national research systems by predatory publishing in a broad comparative perspective. Systematic scrutiny of cross-country differences worldwide is lacking.

This paper helps to fill that gap by examining the propensity to publish in suspected predatory journals for 172 countries in 4 fields of research over the 2015-2017 period. Using the names of journals and publishers on lists by Beall (2016), we derived the ISSNs of 3,293 titles from Ulrichsweb (2016) and searched Scopus (2018a) for them. A total of 324 matched journals with 164 thousand indexed articles was identified. Next, we downloaded from Scopus the number of articles by author's country of origin published in these journals and compared the figures to the total number of indexed articles by country and field. The resulting database provides more representative and comprehensive country-level evidence on publishing in the suspected predatory journals than has been available in any previous study.

Our analysis indicates that there is remarkable heterogeneity in the propensity to publish in suspected predatory journals across countries. In line with earlier evidence, the most affected countries are in Asia and North Africa, but they are not necessarily the same ones cited above. In the most affected countries, including Kazakhstan and Indonesia, around 17 % of articles were published in the suspected predatory journals, while some countries have no articles in this category whatsoever. India's situation also looks daunting, but it is not the worst off. Econometric analysis of cross-country differences shows that countries with large research sectors at the medium level of economic development tend to be most susceptible to predatory publishing. Arab, oil-rich and/or eastern countries are also particularly vulnerable. To the best of our knowledge, this is the first systematic attempt to pin down national research systems at the most risk of falling into the trap of predatory publishing.

No doubt, the lists of predatory, questionable or fake journals are controversial. It should be emphasized that the purpose of this paper is not to evaluate the suspected predatory journals and assess whether they deserve this label or not. Beall (2015 and 2016) developed the identification criteria and put his reputation on the line by curating the lists, which in turn became widely used in empirical research on this topic (see, for instance, Baguess et al. 2019, Bohannon 2013, Bolshete 2018, Cobey et al. 2019, Demir 2018 and 2020, Downes 2020, Erfanmanesh and Pourhossein 2017, Frandsen 2017, Ibba et al. 2017, Kurt 2018, Perlin 2018, Shen and Björk 2015, Shamseer et al. 2017, Wallace and Perri 2018 and Xia et al. 2015). We use this source of data and our competences in comparative research to throw new light on cross-country differences in the propensity to publish in them. In our view, this helps to deepen understanding of the problem of predatory publishing.

The paper proceeds as follows. The second section reviews existing literature on predatory publishing, introduces Beall's lists, and elaborates on their validity and limitations. The third section explains how the dataset has been constructed and how it can be used. The fourth section provides an exploratory analysis of differences across countries and relevant country groups and presents econometric tests of the relationships hypothesized. The concluding section summarizes the key findings and pulls the strands together.

Taking stock of the literature

Predatory publishing

Jeffrey Beall popularized the term *predatory publishing* on his blog (Beall 2016). It is used to describe the practice of abusing paid open-access scientific publishing. In contrast to standard subscription-based models, authors publishing via paid open-access make business directly with publishing houses. They pay article processing fees directly to the publisher of the journal. Both authors and publishers are motivated to publish articles. Predatory journals perform only vague, pro-forma, and in some cases no peer-reviews, and allow publication of pseudoscientific results (Bohannon 2013, Butler 2013). Predatory journals have also been accused of aggressive marketing practices, having fake members of editorial boards and amateur business management (Beall 2015, Cobey et al. 2018, Eriksson and Helgesson 2017a). However, the latter are only side-effects. We use the term *predatory journals* to signify journals suspected of abusing paid open-access to extort fees from authors, and following significantly flawed editorial practices.

The open-access model, though it is a defining element of predatory journals, is not at fault *per se*. The inherent conflict of interest does not have to be exploited. There are effective means to ensure the quality of the editorial practices of journals. Databases dedicated to supporting open-access, such as the Directory of Open Access Journals, are already working to develop operational mechanisms to guarantee quality and to employ transparency measures such as open peer-review, which can easily detect fraudulent publishers. Journals not performing peer-reviews have admittedly nothing to report here. The existence of predatory journals does not mean that the movement calling for democratizing communication of scientific results is fruitless.

Nevertheless, it is challenging to recognize a predatory journal in practice, because there is no clearly defined boundary between journals that follow ethical editorial standards and those that are merely vehicles for exploiting publication fees. Most often, to facilitate awareness and identification, lists are used to identify suspected predatory journals. The most prominent example is Jeffrey Beall's blog (Beall 2016), which was shut down at the beginning of 2017 (Straumsheim 2017).[1] A private company, Cabells, subsequently began to offer a similar list (Silver 2017), the content of which, however, is locked behind a paywall (Cabells 2022). China has announced the formation of a list of 'poor quality' journals (Cyranoski 2018), which was followed by the creation of a list of questionable journals by the National Science Library of the Chinese Academy of Sciences (Zhang et al. 2022), but this list seems to be far more narrow in scope than both of its predecessors.

The inclusion of individual journals on a list should be based on rigid and transparent criteria. Beall (2015) provided a list of criteria that he used to make decisions about journals and publishers. Eriksson and Helgesson (2017a) and Cobey et al. (2018) have also suggested a similar list of characteristics to identify predatory journals. The key set of Beall's criteria points directly to the most salient problem of dubious editorial practices: ("*Evidence exists showing that the publisher does not really conduct a bona fide peer-review"; "No academic information is provided regarding the editor, editorial staff, and/or review board members"*). However, there is also a group of indicators concerning professionalism and/or compliance with ethical standards: ("*The publisher has poorly maintained websites, including dead links, prominent misspellings and grammatical errors on the website"*; "*Use boastful language claiming to be a 'leading publisher' even though the publisher may only be a start-up or a novice organization"*), etc.

Grudniewicz et al. (2019) addressed what they perceived as a lack of agreed definition of predatory publishing by convening dozens of experts on this topic, who arrived at the following consensus: "Predatory journals and publishers are entities that prioritize self-interest at the expense of scholarship and are characterized by false or misleading information, deviation from

---

[1] Anonymous authors continue with Beall's work and regularly update his lists on a new website (Anonymous 2022).

best editorial and publication practices, a lack of transparency, and/or the use of aggressive and indiscriminate solicitation practices" (ibid, pg. 211), which is arguably well in tune with Beall (2015). But when it comes to criteria for identification of predatory journals in practice, they argue for relying on easy to detect defects, such as misinformation, spamming and/or spelling errors, rather than attempting to assess the quality of peer review. By giving up on the latter, however, the identification is likely to miss out predatory journals that have become professionalized and manage to avoid the most obvious blunders, while still neglecting peer review to maximize profits. In this regard, we concur with Moed et al. (2022) that accepting manuscipts without any rigorous form of peer review is the core characteristic of predatory journals that we thus should not leave out.

Kurt (2018) identified 4 pretexts that are often used to justify publication in predatory journals by the authors: i) social identity threat; ii) lack of awareness; iii) high pressure to publish; and iv) lack of research proficiency. The common denominator is urgency. Researchers tend to publish in these journals as a last resort and often refer to institutional pressure, a lack of experience and fear of discrimination from "traditional" journals. Justifications for publishing in predatory journals therefore appears to be a complex mix of factors operating at both personal and institutional levels.

Demir (2018) and Baguess et al. (2019) also argue that the tendency to publish in predatory journals is likely to be related to the quality of research evaluation in the country. In countries where the culture of evaluation and peer pressure pushes researchers to publish in respectable journals, there is little to no motivation to resort to predatory journals, as such behaviour will harm the researcher's reputation. An additional boost to the demand for publishing in fraudulent journals just to clinch points for outputs regardless of merit can occur in systems blindly relying on indexation in bibliometric databases such as Scopus, Web of Science or Medline, which is not accompanied with sufficient evaluation culture.

Predatory publishing can be seen as wasteful of resources. Shen and Björk (2015) estimated the size of the predatory market as high as 74 million USD in 2014, based on article processing fees, and the figure may well have grown significantly since. Perhaps more important than the direct costs, however, are indirect costs stemming from the fact that the opportunity to bypass the standard peer-review process leads researchers astray. Instead of spending their time producing relevant insights, researchers may be increasingly prone to write bogus papers that only pretend to be scientific. If this occurs on an increasing scale, research systems are in peril.

The fact that research published in scientific journals is predominantly funded from public sources only amplifies these concerns.

Beall's lists

Beall (2016) maintained two regularly updated lists of "potential, possible, or probable" predatory journals and publishers, henceforth for the sake of brevity referred to as "suspected predatory": i) a "*list of standalone journals*", which contains suspected individual journals; and ii) a *"list of publishers*", which contains suspected publishers, most of which print multiple journals.

Crawford (2014b) went through every single item on Beall's lists in late March and early April 2014. He found 9,219 journals in total, of which 320 were from the list of standalone journals and 8,899 from the list of publishers. Between 2012 and 2014, about 40 % of those journals published no or fewer than four articles; in other words, they were empty shells, and a further 20 % published only a handful of articles. Another 4 % consisted of dying or dormant journals whose publications fell to a few articles in 2014, and 6 % were unreachable (the web link was broken, for instance). Overall, fewer than 30 % of the identified journals published articles regularly. Fewer than 5 % of the journals appeared "apparently good as they stand", meaning that there was no immediate reason to doubt their credibility, which, however, did not imply that they were in fact credible.

Shamseer et al. (2017) confirmed that Beall's listed journals contained more spelling errors, promoted bogus bibliometric metrics on their websites and their editorial board members were much more difficult to verify than those of 'ordinary' journals. Bohannon (2013) exposed flawed editorial practices by submitting fake scientific articles to journals of publishers from Beall's list. The fake articles were accepted for publication by four-fifths of the journals that completed the review process, which vindicates doubts about their peer review routines. Bagues et al. (2019) showed that journals on Beall's list tend to have low academic impact and cite researchers admitting that editorial practices of these journals are flawed. Journals from these lists truly seem to be doubtful.

Moed et al. (2022) examined journals from Beall's list of publishers with the help of bibliometric analysis using an updated version of database published an earlier version of this paper. First, they found that article output of a random sample of these journals that were not indexed in Scopus had a strong tendency to dwindle and two-fifths of them were discontinued, which confirms that most of them do not succeed to become regular publication venues. Second, they found that the subset of these journals that were indexed in Scopus suffered as a group a strong decline in citation impact and achieved impact levels far below of a control group of other open access journals indexed in Scopus, which they interpret as a signal that in general their scientific relevance is inferior. Finally, however, they also pointed to variability of the suspected predatory journals and that judging by bibliometric records the inclusion of some publishers on Beall's list might be questionable.

Strinzel et al. (2019) compared lists of predatory journals originated by Beall (updated by its anonymous continuator) and more recently launched by Cabells Scholarly Analytics (therein called Cabells) as well as lists of credible journals compiled by the Directory of Open Access Journals (DOAJ) and Cabells using data from December 2018. In terms of journals and publishers indexed, they concluded that there was a considerable overlap between the lists or predatory journals and even speculated that Cabells' list may have used Beall's list as a source, but that there was essentially no or a very limited overlap between them the lists of credible journals.[2] In terms of inclusion criteria, the analysis revealed that both of the lists of predatory

---

[2] Strinzel et al. (2019) remain silent on what explains differences between the lists of predatory journals that they actually identified, but four factors driven by data issues are likely to be in play. First, Beall's lists include only open access journals and publishers, while Cabells' list includes also subscription-based ones. Second, one needs to keep in mind the findings by Crawford (2014b) that most of journals on Beall's lists are empty shells, the findings by Moed et al. (2022) that they tend to dwindle or become discontinued and the findings by Siler et al. (2021) that they become rebranded and morphed into different outlets. The origins of Beall's lists go back to early 2010s and he was well known to focus on adding new records rather than deleting possibly irrelevant old ones, while Cabells' list was launched in 2017, therefore many records that remain indexed in the former might not appear in the latter simply because they ceased to exist or changed in the meantime. Third, Beall's blog was shut down at the beginning of 2017 (Straumsheim 2017) and even though its continuator pledges to update it regularly (Anonymous 2022), it cannot be taken for granted that the updates are as thorough as could have been in Beall's original endeavour, and therefore that new predatory journals that emerged in the meantime may have been recorded in Cabells' list but not in the updated Beall's list. Finally, Strinzel et al. (2019) did not indentify individual journals from Beall's list of publishers, while Cabells' lists of journals and publishers are linked together, hence they compared the restricted list of Beall's standalone only journals with the all-encompassing list of Cabells journals, as the result of which the overlap at the journal level had been underestimated.

journals considered most frequently business practices, including the business model, misinformation on location, spamming and boastful language, but that these aspects were far more dominant for Beall than Cabells, and that the main difference was that Cabells used noticeably more criteria than Beall related to peer review and policy. However, as also acknowledged as a limitation by the authors, the comparison relied only on the number of criteria in each category, not reflecting on their relative weight for indexing, which could have differed significantly.

Limitations

As Eriksson and Helgesson (2017b) state, “the term ‘predatory journal’ hides a wide range of scholarly publishing misconduct.” Some are truly fraudulent, while many others may operate on the margins. However, Beall’s lists force us to work with a binary classification in which a journal and publisher is considered either predatory or not. As Beall did not systematically explain his decisions, it is not possible to make a more detailed quantification of “predatoriness”, though elaborated criteria exist.

Beall’s lists have been strongly criticized for the low transparency of his decision-making process (Berger and Cirasella 2015; Crawford 2014a; Bloudoff-Indelicato 2015). Although the criteria are public, justification of decisions on individual journals and publishers is often not clear and difficult to verify. Beall debated the decisions on his blog or Twitter in some important instances, but very often a journal or publisher was added to the list without justification being provided. The lack of comprehensive, rigid, and formal justification of Beall’s judgments is a major drawback of his list.

In particular, caution is warranted when working with Beall's list of publishers. Classifying an entire publishing house as suspected predatory is a strong judgment, and it cannot be ruled out that some journals which actually apply reputable standards have been listed along the way. The list includes some publishers that maintain broad portfolios of dozens and even hundreds of journals, some of which may not deserve the predatory label, so that using Beall’s list may result in overestimations of true “predators.” It is likely that the overwhelming majority of these journals are of poor quality, but poor quality is not a crime *per se*. One must, therefore, keep in mind that the list of publishers has been painted with a relatively broad brush.

Nevertheless, respectable publishing houses should have zero tolerance for predatory practices. Just as in the banking sector, academic publishing services are based on trust, and if that is lost, the business is doomed. A single journal with predatory inclinations that are not quickly corrected by the publisher can substantially damage the entire brand. Beall's suspected predatory mark signals serious doubts about the publisher's internal quality assurance mechanisms at the very least.

The greatest controversy was triggered by inclusion of the publisher Frontiers on Beall's list of publishers in October 2015. Beall defended this decision by pointing out several articles that, according to him, should not have been published. According to critics of this move, the Frontiers publisher is "legitimate and reputable and does offer proper peer-review" (Bloudoff-Indelicato 2015). Frontiers journals appear to be quite different from typical suspected predatory outlets on the face value of their citation rates. Only 4 journals in Frontiers' portfolio of 29 included in this study are not ranked in the first quartile in at least one field according to the Scimago SJR citation index (Scopus 2018b). Many Frontiers journals are also indexed in the Web of Science and the Directory of Open Access Journals. Hence, judging by the relevance of Frontiers journals for the scientific community, there is a question mark about their inclusion on the Beall's list.

Another concern arises from the timescale. The suspected predatory status used in this study is derived from the content of Beall's lists on 1st April 2016. Jeffrey Beall continuously updated his lists. However, the lists always reflect only current status, with no indication of when the journal and publisher may have become suspected to be predatory. When looking back in time, we may run into the problem of including in the predatory category records that do not deserve that label, because the journal became suspected only a short time before its inclusion to the list. In some cases, older articles published in journals that are currently suspected to be predatory may have gone through a standard peer-review. Hence, historical data must be used with great caution.

Further, Beall's lists are very likely to suffer from English bias. The lists contain mainly journals that at least have English-language websites. In regions in which a large part of scientific output is written in other languages - such as in Latin America, Francophone areas and countries of the former Soviet Union - estimates of the extent of predatory publishing based on Beall's lists may be underestimated, because Beall did not identify suspected predatory journals in local languages. Likewise, Scopus covers scientific literature in English far more comprehensively than publications in other major world languages. This bias should be kept in mind when interpreting cross-country differences.

Database

Our database was built in three steps. First, we compiled a comprehensive overview of journals suspected of predatory practices by matching the lists of standalone journals and publishers by Beall (2016) with records in the Ulrichsweb (2016) database, which provides comprehensive lists of periodicals. Second, we searched the International Standard Serial Numbers (ISSNs) of the journals obtained from Ulrichsweb in Scopus, and downloaded data on authors publishing in these journals by their country of origin. Third, we downloaded the total number of indexed articles by country from Scopus. Ultimately, we obtained not only a full list of suspected predatory journals listed in Scopus but, even more importantly, we also obtained harmonized data on the propensity to publish in these journals by country, which allows us to shed new light on cross-country patterns (for a brief overview of the data generation process see Table 1.1 below).

Beall's lists were downloaded on April 1st, 2016. First, we identified all search terms in each item on the lists. For some entries, Beall presented multiple versions of a journal designation; for example, the journal name and its abbreviation. All available versions were used as a search term. Next, we searched the terms in the Ulrichsweb database for the same day, using an automatic script programmed in Python. When we searched for a standalone journal, the script used the '*title'* field, and for the publisher, the script used the '*publisher*' field. In the end, the algorithm saved all search results. The search request in Ulrichsweb was as follows for standalone journals:

+*(+title:("Academic Exchange Quarterly"))*

and for publishers:

+*(+publisher:("Abhinav"))*

The raw search on Ulrichsweb produced a database of 19,141 results linked to individual entries on Beall's list. Results without ISSNs were removed, as they were most probably not listed in Scopus anyway; this reduced the database to 16,037 search results with 7,568 unique ISSNs. The reduction is due to using multiple search terms related to the same entry and to the ‘fuzziness’ of the Ulrichsweb search.[3] To make sure that the journals are listed by Beall, remaining search results were checked manually. Beall's lists consist of hypertext links, so we compared the ISSN on the journal’s website with the ISSN on Ulrichsweb. If the two ISSNs matched, the entry was retained; if they differed, the entry was removed from our database. A publisher’s identity was confirmed if at least one ISSN listed on its website was found in an entry linked to the publisher's name on Ulrichsweb.

In total, we confirmed 4,665 unique ISSNs associated with Beall's lists. Many journals have dual ISSNs, one for its print and one for its electronic version. The number of individual journals is 3,293, of which 309 featured on the list of standalone journals, 2,952 referred to the list of publishers, and an additional 32 journals appeared on both lists, perhaps because Beall did not recognize that the respective journal was from a publisher already on his list. For simplicity, these journals are considered to belong to the list of publishers.

This is in line with the analysis of Crawford (2014b), which identified fewer than 3,000 journals that published articles regularly, and thus in fact appeared to be continuously in operation. Shen and Björk (2015) found around 8,000 journals that were “active” in the sense that they published at least one article. However, many of these, as per Crawford (2014b), may not publish significantly more than that and are not likely to be registered in databases. Note that there are 1,003 hypertext links on the list of standalone journals, from which it follows that more than two-thirds of these are not included in Ulrichsweb, let alone in more selective databases. Apart from the unverified information on their web pages, there is no information about them. Previous attempts to collect data on suspected predatory journals were far less comprehensive.[4]

[3] The Ulrichsweb search engine uses a ‘fuzzy’ search which does not require perfect matching of strings. For example, when we searched for *Academe Research Journals*, journals of *Academic Research Journals* were also found. This is beneficial because the search is robust to typos, interpunction signs, and small errors written in the search terms. However, it also requires careful manual verification of search results.

[4] For example, Perlin et al. (2018) found only 1100 ISSNs from both the list of publishers and the list of standalone journals using an automatic website crawler and Demir (2018) analyzed only the list of standalone journals.

In the next step, we searched for the presence of these “predatory” ISSNs in the Scopus (2018a) citation database over the period 2015-2017. Once again, this search was performed using an automatic script programmed in Python. The search was performed on March 19th, 2018. For each ISSN detected in Scopus, the script downloaded not only the total number of documents in the “article” category, but also more detailed data on the number of these articles by the author's country of origin. The search request in Scopus was as follows:

ISSN(1234-5678) AND DOCTYPE(ar) AND PUBYEAR > 2014 AND PUBYEAR < 2018

439 ISSNs of 324 individual journals with at least one entry in Scopus were identified, of which 37 appear on the list of standalone journals and 287 on the list of publishers. Thus, nearly 10 % of the journals in our database were indexed in Scopus. In total, 164,073 articles published in these journals were detected, of which 22,235 occur in standalone journals and 141,838 come from the list of publishers, jointly making up 2.8 % of all articles indexed in Scopus during the period under consideration. Hence, the list of publishers, which was rather neglected in previous empirical studies of predatory publishing, is the dominant source. The journals were assigned to four broad fields of research: i) Health Sciences; ii) Life Sciences; iii) Physical Sciences; and iv) Social Sciences, based on the Scopus Source List (Scopus 2018b). If a journal is assigned to multiple fields, it is fully counted in each of them. The database is available for download in Zenodo (Macháček & Srholec, 2022b).

Finally, we obtained data on the total number of articles in Scopus by author's country of origin and field of research over the period 2015-2017, which is the denominator required to compute the penetration of suspected predatory journals in the article output of each country. The download was performed on March 5th, 2020. The search was performed using the following request:

AFFILCOUNTRY(country) AND SUBJAREA(field) AND DOCTYPE(ar) AND PUBYEAR > 2014 AND PUBYEAR < 2018

Table 1.1: Overview of the Data Generating Process

1) Obtaining the ISSNs of suspected predatory journals:

    a. Beall’s lists downloaded on April 1st, 2016.

    b. The names on Beall’s lists were searched for using an automatic script in Ulrichsweb on the same day.

    c. The entries found in Ulrichsweb were manually verified with the help of hypertext links in Beall's lists.

    d. 4,665 ISSNs of 3,295 individual journals were confirmed to be associated with Beall's lists.

2) Searching for “predatory” ISSNs in Scopus:

    a. The “predatory” ISSNs were searched for using an automatic script in Scopus on March 19th, 2018.

    b. 439 ISSNs of 324 individual journals that had at least one entry in Scopus over the period 2015-2017 were identified.

    c. The script downloaded the total number of indexed articles in each journal and the number of these articles by the author's country of origin over the period 2015-2017.

    d. To avoid double-counting articles in journals with ISSN for both print and electronic versions, duplicates were eliminated.

3) Downloading total number of articles in Scopus by country and field of research:

    a. The total number of indexed articles by country over the period 2015-2017 was downloaded using Scopus API on March 19th, 2018.

    b. The total number of indexed articles by country and field of research over the period 2015-2017 was downloaded using Scopus API on March 5th, 2020.

In the Scopus database, an article is fully attributed to a country if affiliation of at least one of its authors is located in that country. Joint articles by authors from different countries are counted repeatedly in each participating country. Hence, the data measure article counts, not fractional assignments. If articles in suspected predatory journals have fewer co-authors than other articles, the predatory articles penetration is underestimated and vice-a-versa; this can be uneven across countries.[5] For some articles, Scopus reports the country of origin as "undefined"; these are excluded from our analysis.[6].

How comes that there are suspected predatory journals in Scopus? Journals need to fulfil a number of selection criteria to become indexed in the database (Scopus 2019). However, these criteria are either formal, such as having an ISSN, online availability and English language abstracts and titles, derived from bibliometrics, such as a minimum threshold of citations, article output and diversity of authors by country, or rely on policies in the sense of what the journal declares to do, for instance, in terms of peer review, rather than what it does in practice. If these boxes are ticked, the journal is very likely to be accepted into Scopus. Yet predatory journals that managed to professionalize their business operation might look like regular scientific outlets on the outside, their bibliometric profile might not differ that much from other fringe journals and they do not shy away from lying about their editorial practices; deception is their defining feature. So this filter is not likely to be effective in keeping out predatory journals that are good pretenders.

Scopus (2021) in a reaction to the earlier publication of this article (Macháček and Srholec 2021) acknowledged that the database included the matched journals that have been identified in our analysis, re-evaluated them and discontinued coverage of more than two-thirds of them: "*All of the 137 suspicious titles mentioned in the paper have gone through the re-evaluation process and as a result for 97 titles (71%), the decision was made to discontinue coverage in Scopus. Also, all other journals listed by Beall that are mentioned in the paper have gone through the re-evaluation process and as a result 65% of these titles were also discontinued*

[5] Unfortunately, the Scopus database does not directly provide harmonized data on the number of authors by country that published in a journal. However, we can count the number of countries, to which at least one author of an article is affiliated, by journal. Based on data for 324 suspected predatory journals and 23,387 other Scopus journals, the average number of country-affiliations turns out to be 1.20 and 1.23, respectively, hence there is not a significant difference and the bias is likely to be rather small. We thank one of the anonymous reviewers for pointing out this potential shortcoming.

[6] Only 1,069 suspected predatory journal articles had an 'undefined' country of origin. Hence, the overwhelming majority of the articles found are included in our analysis.

(ibid, pg. 5)."[7] Moed et al. (2022) also reported that the indexing of about 60% of the suspected predatory journals that they found in Scopus using updated version of our database was discontinued and that 2016 was a peak year in this respect. Scopus thus validated in hindsight that most of these journals were problematic and probably should not have been included in the database in the first place. In the meantime, however, papers published in these journals before they were discontinued remain included in the database possibly misleading by its content unsuspecting readers. [8]

At the same time, it needs to be emphasized that the fact that a matched journal has not been discontinued by Scopus does not signify that it should be absolved from the suspected predatory status. Scopus selection criteria and by extension the re-evaluation criteria, as discussed above, rely heavily on bibliometric indicators and journal's declared policies and only partly check for the attributes that have been proposed to identify predatory journals (Beall 2015, Grudniewicz et al. 2019 and Strinzel et al. 2019), especially with regards to differences between what the journal claims to be the case and reality. Scopus (2021) noted that formal complains that have been raised about publication standards are reflected in the re-evaluation process but does not provide any details, i.e. how many complains are collected, what they are typically about, how they are fact-checked and what represents an offence that is serious enough to discontinue coverage. Until this mechanism of data collection becomes more transparent and widely used by the research community to become representative, one cannot rule out that predatory journals fall through the cracks.

Admittedly, the suspected predatory journals that are indexed in Scopus represent only the tip of the iceberg, which is not representative of the whole business. No matter how imperfect the entry filter of Scopus turns out to be, the journals that made it through represent probably the least ugly part. Nevertheless, from the research evaluation perspective, suspected predatory journals indexed in respected citation databases are more dangerous than ordinary bogus

---

[7] Macháček and Srholec (2021) identified 324 suspected predatory journals with at least one entry in Scopus over the period 2015-2017, which does not correspond with the number of journals cited by Scopus (2021), but the statement makes clear that most of those that were re-evaluated have been discontinued.

[8] An early version of this paper came out in March 2017 (Macháček and Srholec 2017) and was presented at the Scopus Content Selection and Advisory Board Meeting in Prague on November 3, 2017.

journals that few take seriously, because the indexation bestows a badge of quality.[9] All too often, evaluations at various levels rely on this badge and blindly assume that whatever is indexed counts and deserves to be supported by taxpayer's money. Scopus-listed journals are in practice considered 'scientific' by many institutions and even national evaluation systems, such as, for example, in the Czech Republic (Good et al. 2015), Italy (Bagues et al. 2019) and probably many other countries as well. In particular, evaluation systems that do not check the actual content using their own peer-review assessment are most exposed, but such assessment tends to be expensive and difficult to organize, and thus is relatively rare exactly in environments that need this check most.

Cross-country patterns

Out of more than two hundred countries for which the data are available, we excluded dependent territories and countries with fewer than 300,000 inhabitants. The analysis considers evidence from the period between 2015 and 2017, because, as noted above, using older data risks that some of the journals currently featured on Beall's lists were not yet likely to be predatory at an earlier time. However, we use data from three years to increase the robustness of the results. Only countries generating at least 30 articles during this period are included in the analysis. As a result, the final sample consists of 172 countries, which together account for the overwhelming majority of the world's research activity.

The outcome variable used throughout the analysis is the share of articles linked to Beall's lists out of all articles by authors from the given country, hence the share of articles published in suspected predatory journals out of total articles. First, we look at the global picture and examine which countries are most and least affected by predatory publishing. Then, we attempt to pin down the most salient patterns by considering differences between groups of countries. Finally, we investigate how these patterns differ by broad fields of research.

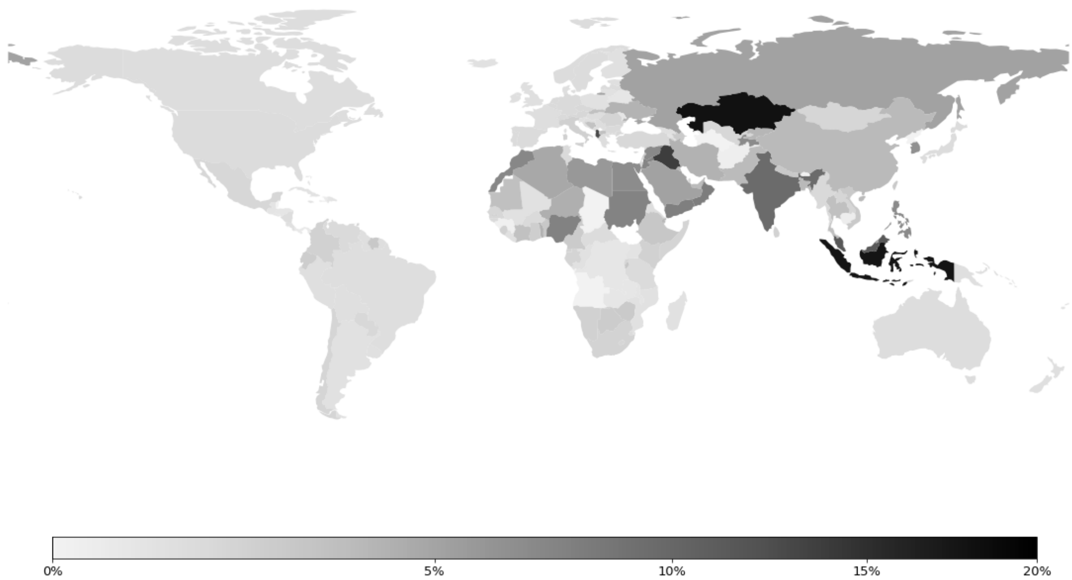
Figure 1.1 displays the results on a world map. The darker the colour, the higher the national propensity to publish in suspected predatory journals. The main pattern is visible at a quick glance; the darkest areas are concentrated in Asia and North Africa. In contrast, Europe, North

---

[9] We use Scopus rather than the Web of Science because it covers substantially more journals (Mongeon and Paul-Hus 2016) and is more vulnerable to suspected predators (Demir 2020; Somoza-Fernández 2016).

and South America and Sub-Saharan Africa are relatively pale. Hence, generally speaking, both the most and least developed countries tend to be relatively less affected, while developing countries with emerging research systems, excepting those in South America, appear to be most in harm's way.

Figure 1.1: % of suspected predatory journal articles in total articles, 2015-2017



Source: Scopus (2018a), author's calculations.

Table 1.2 shows figures for the top and bottom 20 countries. Kazakhstan and Indonesia appear to be the most dire, with roughly every sixth article falling into the suspected predatory category. They are followed by Iraq, Albania and Malaysia, with more than every tenth article appearing in this category. Some of the most severely affected countries are also among the largest in terms of population: India, Indonesia, Nigeria, the Philippines and Egypt, which underlines gravity of the problem. However, small countries that might have been difficult to spot on a world map, such as Albania, Oman, Jordan, Palestine and Tajikistan are also seriously affected. South Korea is by far the worst among advanced countries. All countries on the top 20 list, excepting only Albania, are indeed in or very near Asia and North Africa.

Surprisingly, the opposite end of the spectrum, with the lowest penetration of suspected predatory journal articles, is also dominated by developing countries, including some of even the least developed. In several, for instance Bhutan, Chad and North Korea, there are no authors published in suspected predatory journals whatsoever. This is a rather diverse group of countries scattered across continents. Nevertheless, they have one additional feature in common: most are small countries with underdeveloped research systems. In fact, 13 countries on the bottom 20 list produced fewer than 100 articles per year, on average. It may well be that these research systems are small enough to make direct oversight of the actual content of the manuscripts feasible, in which case, predatory journal articles would have nowhere to hide. In large research systems with thousands of articles produced every year, predatory publishing may more easily fly under the radar of the relevant principals.

Table 1.2: % of suspected predatory journal articles in total articles, top and bottom 20 countries, 2015-2017

Top 20		Bottom 20	
Kazakhstan	17.00	Guatemala	0.74
Indonesia	16.73	Solomon Islands	0.74
Iraq	12.94	Bahamas	0.74
Albania	12.08	Angola	0.72
Malaysia	11.60	Honduras	0.72
India	9.65	Belarus	0.70
Oman	8.25	Congo, Dem. Rep.	0.68
Yemen	7.79	Moldova	0.67
Nigeria	7.31	Afghanistan	0.57
Sudan	7.20	Panama	0.56
Jordan	7.19	Cambodia	0.40
Morocco	6.95	Haiti	0.35
Syria	6.88	Guinea	0.10
Philippines	6.68	Belize	0.00
Egypt	6.65	Bhutan	0.00
Palestine	6.56	Cape Verde	0.00
Tajikistan	6.48	Chad	0.00
South Korea	6.37	Maldives	0.00
Libya	6.06	North Korea	0.00
Brunei	5.44	Turkmenistan	0.00

Source: Scopus (2018a), author's calculations.

Table 1.3 summarizes the main patterns by presenting average propensities to publish in suspected predatory journals by country groups, and provides details by the source list. First, we reiterate the geographical dimension by continents, which confirms that the epicentre of predatory publication is in Asia, while the problem is relatively limited in North and South America. In fact, Suriname, the most affected country in the latter, only ranks 50th in a worldwide comparison. On average, Europe and Africa fall in between the two extremes, but this masks relatively large national differences within these continents along the east-west and north-south axes, respectively. Oceania is also little involved, but there are few countries in the region.[10]

Next, we examine differences by major language zones using indicators obtained from the GeoDist database which measure whether the language (mother tongue, lingua francas or a second language) is spoken by at least 20% of the population of the country (Mayer and Zignago 2011). Only English, French, Spanish and Arabic are recognized separately, as other languages are not spoken in a sufficient number of countries. Note that, in contrast to geography, assignment to language zones is not mutually exclusive, as more than one language can be frequently spoken in the same country. [11]

Admittedly, language zones partly overlap with geography. This is most apparent in South America, which is dominated by Spanish-speaking countries and thus, not surprisingly, the propensities are very similar in both country groups. More revealing is perhaps the fact that Arabic-speaking countries, which are concentrated in North Africa and the Middle East, are the primary hotbeds of predatory publishing. English- and French-speaking countries are far more geographically scattered across the globe.

As noted above, Beall's lists may suffer from English bias. Nevertheless, our results only partially support this expectation. English-speaking countries do not display significantly higher propensities towards suspected predatory publishing than Francophone areas or countries speaking other languages. Spanish-speaking countries turn out to be different, perhaps because we miss predatory journals published in Spanish by relying on Beall's lists and/or Scopus data, but speaking English specifically does not make much difference. Of course, more scholars speak English than do general populations, so tentatively the key take away from these figures should be that, for the most part, language does not seem to be a serious entry barrier into predatory publications.

---

[10] More detailed stratification, such as dividing Asia into South, East, Central and West, or Africa into North and Sub-Saharan, runs into the problem of too few countries in some subgroups, which would make averages unreliable.

[11] For example, there are four countries in which both English and French are spoken by at least 20% of the population (Canada, Cameroon, Israel and Lebanon). Nevertheless, the vast majority of countries are assigned to a single language zone.

Language zones, in turn, reflect broader differences related to religion, culture and history, including past colonial links, which often translate to shared institutions and principles of governance. Arabic countries are likely to appear, on average, highly prone to suspected predatory publishing due to a bundle of these factors that affect how research is organized, evaluated and funded far more than the language itself has an impact. In any case, the language zones are a handy tool to account for broad differences along these lines, especially because such data is available for a very large sample of countries.

Third, it is notable that the top 20 list (Table 1.2) includes oil-rich countries such as Brunei, Iraq, Kazakhstan, Libya, Nigeria and Oman, and a closer look at the data reveals that a few more, including Algeria, Bahrain, Iran, Russia and Saudi Arabia, line up just short of the top 20. Why could there be a connection between oil riches and the susceptibility to predatory publishing at the national level? In countries that benefit from oil-related revenues, fiscal constraints of the governments are eased, so that they can spend on whatever suits them, including the support of academic research, more than other otherwise similar countries. It may not be coincidental that some of the oil-rich countries, particularly in the Middle East but also elsewhere (Sarant 2016, Schmoch 2016), began to invest their resource windfalls in developing indigenous university sectors, while lacking a strong research evaluation culture, which takes time to develop. Although this strategy could be beneficial for long-term development of these countries, if fortunes are perhaps hastily poured into supporting research, there could be undesirable side effects, such as a spike in predatory publishing.

To check whether there is a systematic pattern, we draw on indicators for rents from natural resources in the World Development Indicators database (World Bank 2018), specifically from oil and natural gas, and also for a comparison of rents from other resources, including coal, minerals and forests. Countries are classified as intensive on the respective resources if their resource rents account for more than 5% of GDP; this may sound low, but in practice constitutes a healthy boost to the government budget. The results confirm that countries with an economy intensive on rents from oil and natural gas are on average noticeably more susceptible to suspected predatory publishing than the rest of the world. Moreover, interestingly, this seems to be specific to oil and natural gas, as countries rich in other types of natural resources display even less tendency to this kind of publishing than countries which are not particularly endowed by any of the natural resources considered here.

Fourth, we examine whether there are differences along the level of economic development. For this purpose, we use the World Bank (2016) classification that divides countries into four groups according to gross national income per capita. In line with the anecdotal evidence discussed above, high and low income countries appear to be the least affected.[12] The worst situation is in middle income countries, many of which recognize the role of research for development, and therefore strive to upgrade, but lag significantly behind advanced countries not only in technology, but in their ability to effectively evaluate and govern their emerging research systems. Yet the largest difference in the proclivity to suspected predatory publishing is between lower middle income countries, such as Indonesia, India and the Philippines, and low income countries. Overall, therefore, there seems to be a non-linear, specifically inverse U-shaped, relationship.

Finally, as already mentioned above, the low tendency towards suspected predatory publishing in low income (the least developed) countries may be related to the small size of their public research sectors. To examine whether size matters, we divide the sample into quartiles according to the total number of articles published. Countries with small research sectors do not fall into the most frequent contributors to suspected predatory journals, with the single exception of Tajikistan. In fact, the vast majority rank well below the world average. More than half of low income countries indeed fall into the small size category, and thus it is not surprising that the propensity to suspected predatory publishing proves to be similarly low in both country groups. Again, there seems to be an inverse U-shaped relationship, albeit with a different shape of the distribution.

Next, results are reported by the source list we used to identify predatory journals using three categories: i) Beall's list of standalone journals; ii) Beall's list of publishers excluding Frontiers; and iii) Frontiers. The latter is analyzed separately to account for the controversy surrounding the inclusion of Frontiers on Beall's list of publishers, as already discussed above. Frontiers does exhibit a noticeably different pattern from the other two sources. Authors publishing in

[12] The high income group includes Persian Gulf countries, namely Bahrain, Kuwait, Oman, Qatar, Saudi Arabia and United Arab Emirates, which are rich primarily thanks to oil drilling in the region and in which, except only of Qatar, the propensity to suspected predatory publishing is significantly above the world average. If these countries are excluded, the average propensity to suspected predatory publishing in the high income group drops further down to 1.74%.

Frontiers journals are distributed far more evenly across the country groups and in some respects, such as along income per capita, display even an opposite tendency compared to the other sources lists. On the top 20 list of countries with the highest propensities to publish in Frontiers journals feature Austria, Switzerland, Netherlands, Belgium, Germany or Israel, and in these as well as most other advanced countries Frontiers is the dominant source.[13] As a result, the main patterns identified above are even more pronounced in the total figures excluding Frontiers. From this perspective, Frontiers truly does not look as a typical predatory publisher.

The absolute numbers of articles in suspected predatory journals are also worthy of consideration. In countries with large research systems, predatory publishing can be quite extensive, even if the proportion to total articles does not seem problematic. The main case in point is China, which does not stand out in relative terms with 3.66 % of suspected predatory journal articles in the total national article count, but around 44 thousand articles published in suspected predatory journals had at least one co-author from China; this is by far the largest number worldwide. This means that nearly every fourth suspected predatory journal article has a Chinese co-author. Next are India and the United States, with almost every sixth and ninth suspected predatory journal article co-authored by a researcher from that country, respectively. In these countries, there are legions of researchers who are willing to pay to have their work published in suspected predatory journals.

Table 1.4 provides details on the top 20 most affected countries and the averages across all countries by field of research. The latter indicate that the worldwide propensity to publish in suspected predatory journals is almost two times higher in Social and Life Sciences than in Health and Physical Sciences. Social Sciences are particularly ravaged by this problem in a number of countries: in 7 countries, including the relatively large research systems of Malaysia, Indonesia and Ukraine, more than one fifth of articles appear in suspected predatory journals, and in 14 countries more than one tenth of articles fall into this category. Arguably, the credibility of the whole field is at stake here.

---

[13] Approximately two-thirds of suspected predatory journal articles from advanced countries are published by Frontiers. South Korea is a major outlier among advanced countries, not only because of its high overall penetration of this kind of publishing, but also in the fact that the vast majority of these articles are not in Frontiers journals. Taiwan and Slovakia are similar but to a lesser degree.

Table 1.3: % of suspected predatory journal articles in total articles by country group and source list, 2015-2017

Country group	Number of countries	Total	Source list			Total excl. Frontiers
			Stand-alone	Publishers excl. Frontiers	Frontiers	
Geography:						
Europe	40	1.96	0.32	0.95	0.68	1.27
America	28	1.22	0.10	0.53	0.59	0.63
Asia	49	4.22	0.86	3.01	0.35	3.87
Africa	50	2.33	0.41	1.27	0.64	1.68
Oceania	5	1.14	0.04	0.43	0.67	0.47
Language:						
English spoken	37	2.64	0.41	1.65	0.58	2.06
French spoken	21	2.41	0.35	1.22	0.84	1.57
Spanish spoken	21	1.24	0.11	0.43	0.71	0.53
Arabic spoken	21	5.13	1.17	3.52	0.44	4.69
Other language spoken	86	2.42	0.45	1.49	0.48	1.94
Natural resources rents:						
Oil and natural gas	24	3.90	0.80	2.68	0.41	3.49
Other natural resources	39	1.77	0.23	0.87	0.67	1.10
Other countries	108	2.51	0.45	1.50	0.56	1.95
Income per capita:						
High income	48	2.10	0.22	1.11	0.76	1.33
Upper middle income	44	2.92	0.55	1.95	0.41	2.51
Lower middle income	48	3.28	0.78	2.08	0.42	2.86
Low income	30	1.63	0.16	0.76	0.71	0.92
Size of the research sector:						
Large size	43	2.56	0.35	1.48	0.73	1.83
Medium large size	43	3.49	0.75	2.25	0.49	3.00
Medium small size	43	2.62	0.47	1.69	0.46	2.16
Small size	43	1.59	0.25	0.77	0.58	1.01
All countries	172	2.56	0.46	1.55	0.56	2.00

Source: Scopus (2018a), author's calculations.

Indonesia, Iraq and Oman feature on the top 20 lists in all four fields and Egypt, Iran, Kazakhstan, Libya, Malaysia, Nigeria, Palestine, Sudan and Yemen in three. In these countries, predatory publication practices might have become a systemic problem at the national level, not limited to particular clusters. On the contrary, and perhaps even more interestingly at this point, there are countries in which only specific fields went rogue. For example, China is by far the worst in Health Sciences, but does not appear on any other field list.[14] Albania stands out in

[14] Nevertheless, one must not forget the caveat repeatedly mentioned above that the data predominantly includes journals published in English. China not only has a different language but also its own writing system; thus local problems with the predatory model of publication may largely escape our attention.

Social Sciences only. Likewise, India only looks disreputable in Life and Physical Sciences, Russia in Life and Social Sciences, and Ukraine in Social Sciences.[15]

Table 1.4: % of suspected predatory journal articles in total articles by field of research, top 20 countries 2015-2017

Health Sciences		Life Sciences		Physical Sciences		Social Sciences	
China	11.72	Kazakhstan	28.10	Indonesia	22.31	Albania	37.04
Libya	6.20	Iraq	16.55	Malaysia	11.77	Malaysia	29.15
Taiwan	4.87	Syria	14.29	Philippines	10.90	Yemen	28.89
Egypt	4.84	India	13.59	Iraq	10.66	Indonesia	27.21
South Korea	4.73	Algeria	10.99	Jordan	9.19	Tajikistan	25.64
Algeria	4.58	Egypt	10.94	India	8.65	Ukraine	22.63
Luxembourg	4.57	Togo	10.37	Yemen	8.36	Kazakhstan	21.78
Suriname	4.55	Palestine	10.09	Sudan	8.05	Russia	17.54
Saudi Arabia	4.54	Libya	9.39	Morocco	7.86	Brunei	12.60
Nigeria	4.48	Indonesia	9.11	Oman	7.70	Oman	12.39
Iraq	4.36	Nigeria	9.10	South Korea	7.54	Iraq	12.24
Palestine	4.13	Oman	8.77	Kazakhstan	7.17	Azerbaijan	12.15
Indonesia	4.05	Morocco	8.42	Bahrain	6.70	Iran	11.32
Sudan	4.01	Sudan	7.91	Liberia	6.45	Syria	10.11
Iran	3.83	Iran	6.93	Palestine	6.31	Thailand	9.94
Malaysia	3.79	Russia	6.61	Nigeria	6.31	Nigeria	9.28
Chile	3.76	Yemen	6.49	Brunei	5.96	Slovakia	9.27
Italy	3.63	Macedonia	6.19	Egypt	4.99	Bahrain	9.04
UAE	3.62	Niger	6.02	Saudi Arabia	4.85	Jordan	8.13
Oman	3.56	Mauritania	6.00	Libya	4.62	Kyrgyzstan	8.06
All countries	1.98	All countries	3.39	All countries	1.96	All countries	3.99

Note: Journals can be assigned to multiple fields of research. Only countries with at least 30 total articles in the respective field of research. Source: Scopus (2018a), author's calculations.

Overall, we have identified a handful of factors which seem to be relevant for explaining cross-country differences in the propensity to suspected predatory publishing, and which beg for more elaborate examination. Nevertheless, tabulations of the data can only get us so far in isolating their individual effects. Due to limited space and because a combination of several factors appears to be in play, we do not delve deeper into descriptive evidence by field of research, but rather explore these patterns using a multivariate regression framework in the next section. The

[15] In general, there are far more former socialist countries, especially former members of the Soviet Union, on the top 20 list in Social Sciences than in other fields. Social Sciences were particularly isolated, indoctrinated and devastated during the communist era, so it is not surprising that this is the case.

full results at the country-level in total and by field of science are available for download as available for download in Zenodo (Macháček & Srholec, 2022b). .[16][17]

Regression analysis

In this section, we explore the cross-country differences with the help of an econometric model. The main focus of the analysis is on testing the hypothesized relationships between the level of economic development measured by GDP per capita, size of the (public) research sector measured by the total number of articles and the propensity to suspected predatory publishing, while controlling for other relevant factors. The empirical model to be estimated is as follows:

(1) $Y_{ij} = \alpha + \beta\, GDP_i + \gamma\, SIZE_i + \delta\, X_i + \mu_j + \varepsilon_c$

where the outcome variable *Y* is the proportion of articles published in suspected predatory journals, variously defined, *GDP* per capita represents the level of economic development, *SIZE* represents the size of the research sector, *X* is the set of country-level control variables, μ is a fixed effect for the field of research represented by respective dummies, *i* denotes a country, *j* denotes a field of research and ε is the standard error term. Hence, the basic unit of analysis is a field of research in a given country. Since differences between fields of research are fully accounted for by the fixed effects, the estimated coefficients of the country-level variables explain exclusively within fields variability.

The dependent variable is a proportion that falls between zero and one. The Ordinary Least Squares (OLS) estimator tends to produce predicted values outside of this range and assumes linear relationships. Both problems are addressed by using a fractional logit (binomial) in the Generalized Linear Models (GLM) framework. Robust standard errors derived from Huber-White sandwich estimators are reported. Only observations with at least 30 total articles in the respective country-field and with full data available for the explanatory variables are included

[17] Note that most of the patterns by country groups identified in the total data also apply by field of research, as also vindicated by the regression results below.

in the estimation sample. As the result, the econometric analysis is limited to 630 observations in 163 countries.[18] All estimates are performed in Stata/MP 15.1.

Whenever possible we use continuous variables to measure the explanatory factors, as though the number of observations is essentially quadrupled by using the field specific data, the sample is still relatively small. As envisaged above, GDP per capita (PPP, constant 2011 international dollars) is used to measure the level of economic development and the total number of articles indexed in Scopus is used as a rough proxy for the size of the research sector. Oil and natural gas rents (% of GDP) are used to control for the availability of extra resources that ease the fiscal constraints of the governments to invest in research. Latitude and longitude of the country's centroid, instead of plain continental dummies, are used to account for geography. However, the only way to control for the language zones is to use dummies. GDP per capita and the size of research sector variables are used in logs to curtail the impact of outliers. All variables refer to (if applicable averages over) the reference period 2015-2017. For descriptive statistics, definitions and sources of the variables entering the regression analysis, see Appendix Tables Appendix Table 1.6: Descriptive statistics of the variables, 2015-2017 and Appendix Table 1.7: Definition and sources of the variables.

The regression analysis is used as a descriptive tool in this paper. The purpose of the regression model is to test whether the broad cross-country patterns identified above hold in a multivariate framework, when the possible influence of other relevant factors is accounted for. It should be emphasized that the cross-sectional nature of the data does not allow for testing of causality, the estimated relationships indicate correlations, and the results should therefore be interpreted with caution.

Table 1.5 provides results for the benchmark outcome variable of total suspected predatory publishing (Column 1), then results are replicated separately by the source list (Columns 2-4) and finally estimated for the total, excluding Frontiers (Column 5). Since the descriptive overview revealed that there could be a non-linear relationship between the propensity to suspected predatory publishing on the one hand and the level of economic development as well

---

[18] Cuba, Eritrea, North Korea, Somalia and Syria are excluded due to missing data on GDP per capita. Comoros, Djibouti, Timor-Leste and Turkmenistan are eliminated because they did not generate more than 30 total articles in any of the fields of research.

as the size of the research sector on the other hand, we test for this possibility by including the respective variables in squared terms.

GDP per capita has a significantly positive main effect, but the negative squared term indicates that there is indeed an inverse U-shaped relationship. The results confirm that the proclivity to suspected predatory publishing has a tendency to increase with the level of economic development, but only up to a point, after which the relationship turns negative. Hence, countries at a medium level of development are the most vulnerable. Likewise, the size of the research sector comes out with a significantly positive main effect and a negative squared term, thus the same interpretation applies, albeit the relationship is estimated to be far less curvilinear.[19]

Some of the control variables prove to have even more statistically significant coefficients. First, more reliance on oil and natural gas rents, which ceteris paribus loosens the fiscal constraints of governments, is strongly positively associated with suspected predatory publishing. Of course, this is not to say that such resources should not be used to fund research, but there is a catch. Second, Arabic countries are confirmed to be particularly prone to suspected predatory publishing, even after oil and natural gas rents and other factors are accounted for, so there is something special about this area. Further, English is assumed to primarily control for the suspected language bias of Beall's lists and Scopus, but this worry is not supported by the results. Finally, longitude has a significantly positive coefficient, so farther east of the Greenwich meridian implies higher inclinations towards suspected predatory publication.

As far as the comparison by source list is concerned, the results confirm that Frontiers has a different modus operandi than the rest of the pack. If only articles in Frontiers journals are considered, for instance, GDP per capita has statistically significant but opposite signs from the benchmark results. In fact, the model explains this outcome variable quite poorly, from which follows that a different approach is needed to get to bottom of what is up with this publisher. Although there is no evidence in the data presented upon which we can judge whether the inclusion of Frontiers on Beall's list was justified or not, the results at the very least clearly

[19] If the squared terms are excluded from the model, both coefficients come out highly statistically significant, but GDP per capita has a negative sign while the size of research sector has a positive sign.

indicate that Frontiers is atypical. Henceforth, therefore, we focus on the outcomes excluding Frontiers.[20]

Table 1.5 Explaining propensity to suspected predatory publishing, GLM with logit link for binomial family, 2015-2017

	(1) Total	(2) Standalone	(3) Publishers excl. Frontiers	(4) Frontiers	(5) Total excl. Frontiers
Constant	-6.405***	-11.227***	-7.690***	-5.991***	-7.936***
	(0.877)	(1.941)	(1.393)	(0.778)	(1.270)
GDP per capita	0.308*	0.838***	0.450	-0.301*	0.535**
	(0.182)	(0.284)	(0.285)	(0.158)	(0.255)
GDP per capita squared	-0.100***	-0.296***	-0.149***	0.113***	-0.180***
	(0.034)	(0.068)	(0.054)	(0.027)	(0.049)
Size of the research sector	0.405**	1.042**	0.446	0.174	0.588**
	(0.188)	(0.408)	(0.298)	(0.178)	(0.272)
Size of the research sector squared	-0.017*	-0.050**	-0.019	-0.005	-0.027*
	(0.010)	(0.021)	(0.016)	(0.009)	(0.015)
Oil and natural gas	0.019***	0.027***	0.023***	-0.011*	0.024***
	(0.005)	(0.007)	(0.007)	(0.006)	(0.007)
English spoken	-0.095	-0.171	-0.190	0.022	-0.183
	(0.115)	(0.179)	(0.178)	(0.114)	(0.157)
French spoken	-0.088	-0.321	-0.179	0.245**	-0.215
	(0.119)	(0.234)	(0.178)	(0.106)	(0.173)
Spanish spoken	-0.145	-0.544	-0.481	0.246	-0.498*
	(0.188)	(0.408)	(0.323)	(0.180)	(0.280)
Arabic spoken	0.532***	0.648**	0.681***	0.102	0.686***
	(0.175)	(0.258)	(0.215)	(0.124)	(0.209)
Latitude	0.003	0.013**	0.001	-0.001	0.003
	(0.003)	(0.006)	(0.004)	(0.002)	(0.001)
Longitude	0.005***	0.005***	0.008***	-0.001	0.007***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Field of research	Included	Included	Included	Included	Included
AIC	153.03	60.38	108.63	70.63	125.86
BIC	219.72	127.07	175.32	137.32	192.55
Number of research fields	4	4	4	4	4
Number of countries	163	163	163	163	163
Number of observations	630	630	630	630	630

*Note: Only countries with at least 30 total articles in the respective field of research. The dependent variable is the proportion of suspected predatory journal articles in total articles. Robust standard errors are in parentheses. *, **, *** denote significance at the 10, 5 and 1 per cent levels.*

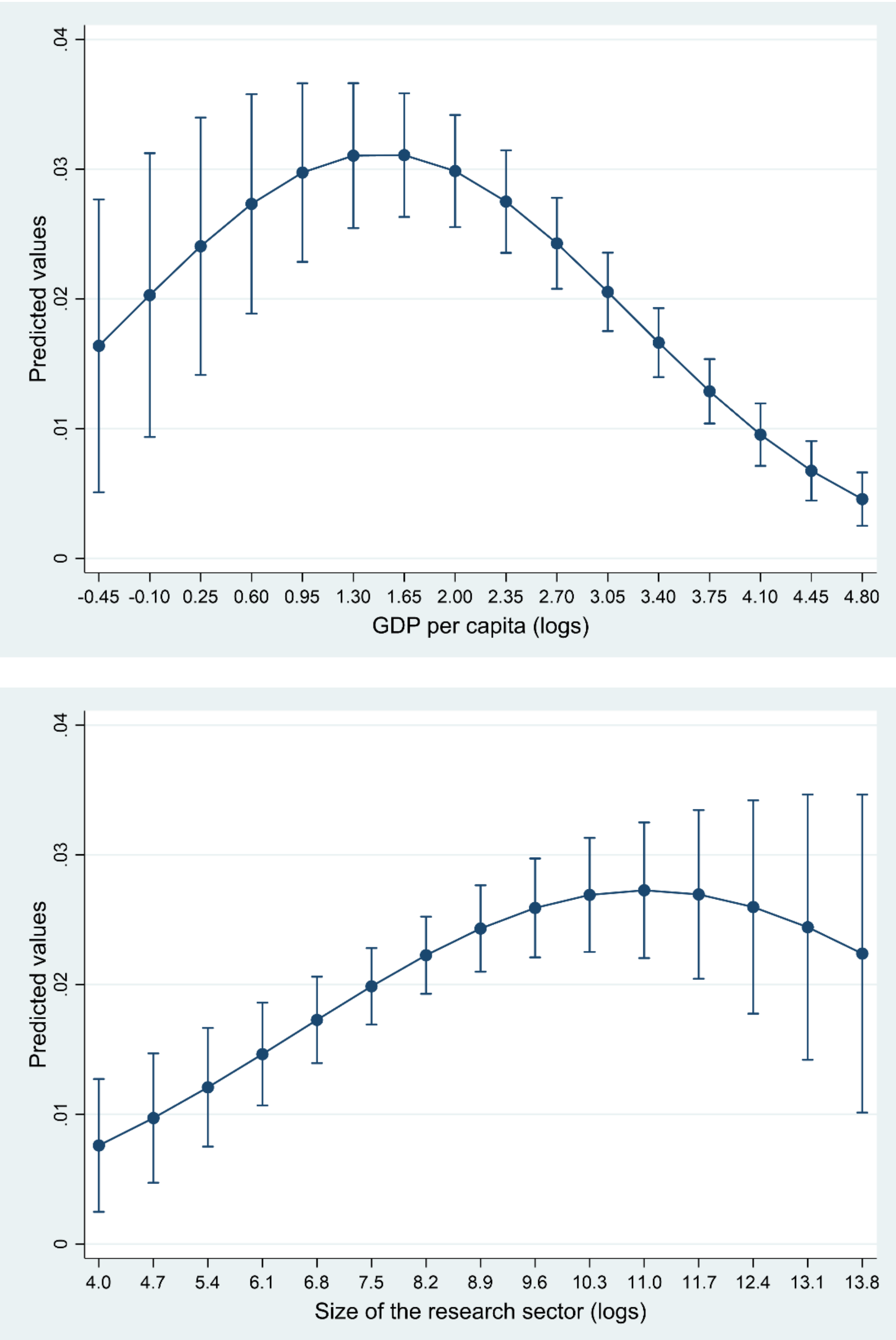
[20] It needs to be emphasized that the authors of this article have never had any connection to the Frontiers or any of their journals in any capacity.

Figure 1.2 gives graphical representations of the estimated relationships of main interest, which provide a handy platform for discussing the results in more detail. The figures clearly illustrate that these relationships follow an inverse U-shaped curve. The propensity to suspected predatory publishing increases with GDP per capita up to approximately the level of countries like India, Nigeria and Pakistan after which, however, there is a steep decline. Along the size measure there is initially a steady increase of suspected predatory publishing until a turning point at the level of countries with relatively large research systems like Malaysia and Saudi Arabia, which is followed by only a slight decrease for the largest ones. The overlapping confidence intervals indicate that, for GDP per capita, the relationship differs most significantly between medium and highly developed countries, while for the size measure the difference is mainly between small and medium research sectors. So what does this mean?

GDP per capita is used for a lack of better measurements that are more intimately related to how a research system is organized and that would be available for a broad sample of countries, including many developing ones. Nevertheless, GDP per capita tends to be highly correlated to many other salient measures. What is likely to make the key difference between medium and high developed countries that drives the results presented in this study is capability to perform meaningful research evaluation, including advanced scientometrics and peer-review of actual content of published papers, that does not fall back on only counting the number of articles indexed in Scopus or elsewhere, regardless of quality and merit. If the government is not able to set the right mix of incentives to the public research sector, which is arguably very difficult even in advanced countries, those who do not shy away from predatory publishing have free rein.

Size is an important consideration, as noted above, because large research systems are more complex and therefore notoriously more difficult for governments to evaluate, manage and steer than small systems. If two countries maintain equally primitive research evaluation frameworks, one with a large research sector composed of dozens of diverse institutions will tend to be more susceptible to predatory publishing than one with a tiny research sector composed of perhaps only a few easy-to-oversee workplaces. Large research systems suffer from a certain degree of anonymity, blind spots and dark corners, in which predatory publishing flourish. Around the turning point, however, the system becomes large enough to warrant investment in advanced research evaluation capabilities, which make life more difficult for those exploiting the loopholes, so that the relationship between predatory publishing and size flattens and even curves slightly down.

Figure 1.2 Estimated effects of GDP per capita (upper figure) and size of the research sector (lower figure) on the propensity to suspected predatory publishing (total excluding Frontiers), GLM with logit link for binomial family, 2015-2017



Note: Based on results in Column 5 of Table 1.5. Predictive margins with 90% confidence intervals are displayed.

Conclusions

Taken at face value, the evidence presented in this paper indicates that countries at a medium level of economic development and with large research sectors are most prone to publishing in suspected predatory journals. This should be a dire warning for developing countries which devote large resources to support research, but which may not pay sufficient attention to upgrading their research governance capabilities, including research evaluation framework. Moreover, the evidence suggests that oil-rich and/or Arabic and/or eastern countries tend to be particularly vulnerable, which completes the picture of who should be primarily on the lookout for predators.

Nevertheless, the general patterns are from a bird's-eye view, so there are exceptions driven by idiosyncratic factors. The prime example of an outlier appears to be Albania, which does not feature most of the high-risk characteristics, but still is among the most affected countries. Predatory publishing is a truly global phenomenon, from which no emerging research system is entirely safe. Policymakers in developing countries that do not fit the description of the main risk group should not be fooled into thinking that the problem does not concern them, because if they flinch in their vigilance, their homeland may end up on the list of the most affected countries next time.

The results are broadly in line with previous estimates by Shen and Björk (2015), Xia et al. (2015), Demir (2018), as well as Wallace et al. (2018), in the sense that Asia and North Africa provide the most fertile grounds for predatory publishing and that in particular India and Nigeria belong to the main sources. However, this paper not only gathered one of the most comprehensive databases of suspected predatory journals, and used far more complete evidence than previous studies, but also provided a much higher level of granularity on the cross-country differences. In fact, a number of countries not mentioned in previous studies are shown here to be likely to suffer substantially from the problem of predatory publishing. In addition, this paper is the first to study the cross-country differences systematically in an econometric framework.

A major limitation of this study is that we can only speculate that the way in which research is evaluated in each country makes the primary difference, whether this includes research organizations at the national level, projects supported by funding agencies, and/or even individuals working on career progression. Ideally, we would like to take characteristics of the research evaluation framework directly into account, including whether evaluation primarily concerns quantity or quality, whether formulae based on quantitative metrics is used, how advanced the underlying bibliometric approach is, whether insights from peer review assessment are factored in, and, consequently, what principles are applied when allocating research funding. Unfortunately, indicators of this kind are not available for more than a handful of advanced countries, which are not the most relevant here. To pin down the impact of these factors on the propensity to predatory publishing remains an important challenge for future research on this topic.

Better insight on seniority of researchers publishing in predatory journals could help target better policies tackling predatory publishing. If the majority of authors were young and inexperienced then campaigns such as *Think. Check. Submit.* could improve the awareness among the target group. However, if the structure of authors spans from doctoral students to senior researchers with rich publication history, then such campaign may not be efficient as it is more likely that the incentive to publish in these journals is built within the researchers' evaluation scheme. The data available for this paper does not allow such analysis, but in the future the analysis may be extended to bibliometric version of Scopus database (such as the CWTS database system) and control for the researchers' academic age or number of papers published.

Another limitation is the cross-sectional nature of the analysis that, as explained above, stems from the fact that historical data is not reliable. Longitudinal data would allow for more elaborate tests, particularly with respect to causality, than those employed in this paper. There are also likely to be lags in the cause–effect relationships that could be detected when long time series become available. In any case, the three-year period studied here is rather short, as predatory publishing is a relatively recent and fairly dynamic phenomenon. This may have influenced the results and the list of most affected countries may look somewhat different if a similar exercise is repeated in a few years, which would be desirable.

It should be stressed that the results of this paper should not be interpreted to mean that developing countries should invest less in research, because this would undermine their emerging and often fragile national innovation systems and ultimately thwart productivity growth (Fagerberg and Srholec 2009). However, it is fair to issue a cautionary note that predatory publishing has the potential to complicate research evaluation and therefore effective allocation of research funding greatly in many corners of the world. Developing countries aiming to embark on a technological catch-up trajectory need to take these intricacies more seriously than ever.

Scopus needs to stay focused on discontinuing coverage of questionable journals and most importantly step-up efforts to prevent their indexing in the first place, as once they are allowed in, the content they publish before their potential discontinuation remains in the database forever. Scopus should strive to find a way to fact-check whether the journal adheres to the declared editorial practices, most prominently how the peer-review process is performed in practice; possibly engaging the research community more actively with regards to collecting data on complains about publication standards in both currently and prospectively indexed journals. Unless the selection criteria are upgraded to reflect not only the declared policy but also reality on the ground and/or the bar for inclusion in terms of bibliometric criteria is raised significantly, new questionable journals will keep creeping in the database, including rebranded and transformed business operations that have been flagged as predatory in the past. In the meantime, evaluators, research managers or university rankings that use Scopus data as inputs in their decisions need to be mindful about it.

Last, but not least, as already discussed above, Beall's lists no doubt have limitations. Beside the lack of transparency of the decisions to list some of the journals and publishers, the most obvious one has become the fact that Jeffrey Beall stopped curating the lists under his name at the beginning of 2017 (Straumsheim 2017), as the result of which the data has become gradually outdated. Even though his lists continue to be maintained by someone on a new website (Anonymous 2022), their updates have not been authorized, which makes their use problematic for research purposes. Future research on more recent evidence on this topic ought to look for different data sources. Cabells Predatory Reports that have been developed in the meantime seem promising (Cabells 2022), but this list is proprietary, locked behind a paywall and the database as a whole is not easily available.[21] Clearly, the research community needs to continue

[21] For example, in July 2021, we discussed with Cabells the possibility to obtain access to early version of the list that underlies their Predatory Reports for the purpose of running a replication study with regards to evidence presented in this paper and comparing the results. After a detailed discussion, however, Cabells concluded that this would not be desirable to do and the access has not been granted to us.

efforts to improve identification, measurement and understanding of the problem of predatory publishing.

References

Abramo, G., Aguillo, I. F., Aksnes, D. W., Boyack, K., Burrell, Q. L., Campanario, J. M., ... & Waltman, L. (2022). Retraction of Predatory publishing in Scopus: Evidence on cross-country differences lacks justification. Scientometrics, 1-3.

Anonymous (2022). Potential predatory scholarly open-access publishers. https://beallslist.net [cited 24 June 2022].

Baguess M., Sylos-Labini M., Zinovyeva N. (2019). A Walk on the Wild Side: `Predatory' journals and information asymmetries in scientific evaluations. Research Policy. 48(2), 462-477

Beall, J. (2013). Predatory publishing is just one of the consequences of gold open access. Learned Publishing. 26(2), 79-84.

Beall, J. (2015). Criteria for Determining Predatory Open-Access Publishers. Available from https://beallslist.weebly.com/uploads/3/0/9/5/30958339/criteria-2015.pdf [cited 19 May 2018].

Beall, J. (2016). Scholarly Open Access: Critical analysis of scholarly open-access publishing (Beall's blog) Available from https://scholarlyoa.com; shutdown January 2018, archived at https://archive.org/web/; [cited 1 April 2016].

Berger, M., Cirasella, J. (2015). Beyond Beall's List Better understanding predatory publishers. College & Research Libraries News. 76, 132-135.

Bloudoff-Indelicato, M. (2015). Backlash after Frontiers journals added to list of questionable publishers. Nature, 526, 613.

Bohannon J. (2013). Who's afraid of peer-review. Science, 342 (6154). DOI: 10.1126/science.342.6154.60.

Bolshete, P. (2018). Analysis of thirteen predatory publishers: A trap for eager-to-publish researchers. Current Medical Research and Opinion, 34, 157-162.

Butler, D. (2013). The dark side of publishing. Nature, 495(7442), 433.

Cabells (2022). Predatory Reports, Cabells Scholarly Analytics, https://www2.cabells.com/about-predatory, [cited 29 June 2022]

Cobey K. D., Grudniewicz A., Lalu, M.M., Rice D.B., Raffoul H., Moher D. (2019). Knowledge and motivations of researchers publishing in presumed predatory journals: a survey. BMJ Open, 9(3), e026516.

Cobey, K. D., Lalu, M. M., Skidmore, B., Ahmadzai, N., Grudniewicz, A., & Moher, D. (2018). What is a predatory journal? A scoping review. F1000Research, 7:1001.

Crawford, W.(2014a). Ethics and Access 1: The Sad Case of Jeffrey Beall. Cites & Insights. 14(4). 1–14.

Crawford, W. (2014b). Journals, „Journals“ and Wannabes: Investigating The List. Cites & Insights. 14(7): 1–24.

Cyranoski, D. (2018). China awaits controversial blacklist of 'poor quality' journals. Nature News. Available from https://www.nature.com/articles/d41586-018-07025-5 [cited 19 May 2018].

Demir, S. B. (2018). Predatory journals: Who publishes in them and why? Journal of Informetrics, 12(4), 1296–1311. doi:10.1016/j.joi.2018.10.008

Demir, S. B. (2020). Scholarly databases under scrutiny. Journal of Librarianship and Information Science, 52, 150-160.

Downes, M. (2020). Thousands of Australian academics on the editorial boards of journals run by predatory publishers. Learned Publishing, 33, 287-295.

Eriksson, S., Helgesson, G. (2017a). The false academy: predatory publishing in science and bioethics. Medicine, Health Care and Philosophy. 2017;20(2): 163-170.

Eriksson, S., Helgesson, G. (2017b). Time to stop talking about "predatory journals." Learned Publishing, 31(2), 181–183. doi:10.1002/leap.1135.

Erfanmanesh, M. and Pourhossein, R. (2017). Publishing in predatory open access journals: A case of Iran. Publishing Research Quarterly, 33, 433-444.

Fagerberg, J. and Srholec, M. (2009). Innovation systems, technology and development: Unpacking the relationship(s). In Lundvall, B.-A., Joseph, K.J., Chaminade, C. and Vang, J. (Ed.), Handbook of innovation systems and developing countries. Building domestic capabilities in a global context, Cheltenham, Edward Elgar, 83-115.

Frandsen, T. F. (2017). Are predatory journals undermining the credibility of science? A bibliometric analysis of citers. Scientometrics, 113, 1513-1528.

Gallup, J. L., Sachs, J. D., Mellinger, A. (1999). Geography and Economic Development. Harvard University. CID Working Paper no. 1/1999.

Good B, Vermeulen N, Tiefenthaler B, Arnold E. (2015). Counting quality? The Czech performance-based research funding system. Research Evaluation, 24, 91-105.

Grudniewicz, A., Moher, D., Cobey, K. D. (2019). Predatory journals: no definition, no defence. Nature, 576, 210-212.

Ibba, S., Pani, F.E., Stockton, J.G., Barabino, G., Marchesi, M., Tigano, D. (2017). Incidence of predatory journals in computer science literature. Library Review, 66, 505-522.

Kurt, S. (2018). Why do authors publish in predatory journals?. Learned Publishing, 31(2), 141-147.

Macháček, V. and Srholec, M. (2017). Predatory journal in Scopus. IDEA study 2 / 2017, https://idea-en.cerge-ei.cz/files/IDEA_Study_2_2017_Predatory_journals_in_Scopus/mobile/index.html#p=1.

Macháček, V., Srholec M. (2021). Predatory publishing in Scopus: evidence on cross-country differences. Scientometrics, 126, 1897–1921.

Macháček, V., Srholec, M. (2022a). Retraction Note to: Predatory publishing in Scopus: evidence on cross-country differences. Scientometrics, 127, 1667.

Macháček, V., & Srholec, M. (2022b). Predatory publishing in Scopus: Evidence on cross-country differences. Zenodo. https://doi.org/10.5281/zenodo.7006761

Mayer, T. Zignago, S. (2011). Notes on CEPII's distances measures: The GeoDist database. CEP II, Working Paper No 2011 – 25, http://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877.

Moed, F., H., Lopez-Illescas, C., Guerrero-Bote, V. P. and de Moya-Anegon, F. (2022). Journals in Beall's list perform as a group less well than other open access journals indexed in Scopus but reveal large differences among publishers. Learned Publishing, 35, 130–139.

Mongeon, P., Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213-228.

Perlin, M. S., Imasato, T., & Borenstein, D. (2018). Is predatory publishing a real threat? Evidence from a large database study. Scientometrics, 116(1), 255-273.

Retraction Watch (2021). Authors object after Springer Nature journal cedes to publisher Frontiers' demand for retraction. https://retractionwatch.com/2021/09/07/authors-object-after-springer-nature-journal-cedes-to-publisher-frontiers-demand-for-retraction/?fbclid=IwAR3C6QDUQCcLQrhKbKJdunleVb99QP-jEbm56299PgZNXrESHsGZMTclSxw [cited 14 June 2022].

Sarant, L. (2016). The Middle East: An end to oil dependency. Nature, 537, S6–S7.

Scopus (2018a). Scopus on-line database. https://www.scopus.com [cited 19 March 2018].

Scopus (2018b). Scopus Source List (May 2018 version). Current version is available at https://www.scopus.com/sources [cited 21 May 2019].

Scopus (2019). Content Policy and Selection; Available from: https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection; [cited 19 May 2019].

Scopus (2021). The importance of high-quality content: curation and reevaluation in Scopus. Available from: https://www.elsevier.com/__data/assets/pdf_file/0004/891058/ACADLIBSCARTImportance-of-high-quality-contentWEB.pdf [cited 13 June 2022]

Shen C, Björk, B-C. (2015). Predatory' open access: a longitudinal study of article volumes and market characteristics. BMC Medicine. 13(230), 1-15.

Schmoch, U., Fardoun, H. M., and Mashat, A. S. (2016). Establishing a World-Class University in Saudi Arabia: intended and unintended effects. Scientometrics, 109, 1191–207.

Shamseer, L., Moher, D., Maduekwe, O., Turner, L., Barbour, V., Burch, R., Clark, J., Galipeau, J., Roberts, J., Shea, B. J. (2017). Potential predatory and legitimate biomedical journals: can you tell the difference? A cross-sectional comparison. BMC Medicine, 15(28).

Siler, K., Vincent-Lamarre, P., Sugimoto, C. R. and Larivière, V. (2021). Predatory publishers' latest scam: bootlegged and rebranded papers. Nature, 598, 563-565.

Silver, A. (2017). Pay-to-view blacklist of predatory journals set to launch. Nature News. Available from https://www.nature.com/news/pay-to-view-blacklist-of-predatory-journals-set-to-launch-1.22090 [cited 19 May 2019].

Somoza-Fernández, M., Rodríguez-Gairín, J. M., & Urbano, C. (2016). Presence of alleged predatory journals in bibliographic databases: Analysis of Beall's list. El profesional de la información, 25(5), 730.

Srholec (2021). RETRACTION REBUKE: Predatory publishing in Scopus: evidence on cross-country differences. https://inovacnipolitika.blogspot.com/2021/09/retraction-rebuke-predatory-publishing.html [cited 14 June 2022].

Straumsheim, C. (2017). No More 'Beall's List'. Inside Higher Ed. Available from: https://www.insidehighered.com/news/2017/01/18/librarians-list-predatory-journals-reportedly-removed-due-threats-and-politics; [cited 27 March 2017].

Strinzel, M., Severin, A,, Milzow, K., Egger, M. (2019). Blacklists and Whitelists To Tackle Predatory Publishing: a Cross-Sectional Comparison and Thematic Analysis. mBio, 10, 10:e00411-19.

Ulrichsweb (2016). Ulrichsweb – Global Serials Directory. 2019; Available from: http://ulrichsweb.serialssolutions.com/; [cited 1 April 2016].

Wallace, F. H., Perri, T. J. (2018). Economists behaving badly: publications in predatory journals. Scientometrics, 115, 749-766.

World Bank (2016). How does the World Bank classify countries? Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries; [cited 10 October 2016].

World Bank (2018). World Development Indicators (last updated July 2018), New York, World Bank.

Xia, J., Harmon, J. L., Connolly, K. G., Donelly, R. M., Anderson, M. R., Howard, H. A.(2015). Who Publishes in "Predatory" Journals? Journal of the Association for Information Science and Technology, 66(7): 1406-1417.

Zhang, L., Wei, Y, Sivertsen, G., Huang, Y. (2022). The motivations and criteria behind China's list of questionable journals. Learned Publishing, https://doi.org/10.1002/leap.1456

Appendix

Supplementary information 1.1: A list of predatory journals in Scopus file

22_QSS_MachacekSrholec_Supp1.xlsx

Supplementary information 1.2: Country-level data file

22_QSS_MachacekSrholec_Supp2.xlsx

Appendix Table 1.6: Descriptive statistics of the variables, 2015-2017

	Mean	St. dev.	Min	Max	N
Dependent variables:					
Total	0.028	0.039	0	0.370	630
Standalone	0.005	0.013	0	0.216	630
Publishers excl. Frontiers	0.016	0.033	0	0.370	630
Frontiers	0.007	0.009	0	0.057	630
Total excl. Frontiers	0.021	0.039	0	0.370	630
Explanatory variables:					
GDP per capita	2.341	1.211	-0.443	4.773	163
Size of the research sector	8.355	2.323	3.989	14.071	163
Oil and natural gas	2.676	7.124	0	48.318	163
English spoken	0.221	0.416	0	1	163
French spoken	0.129	0.336	0	1	163
Spanish spoken	0.123	0.329	0	1	163
Arabic spoken	0.117	0.322	0	1	163
Latitude	20.454	24.752	-41.814	67.470	163
Longitude	20.439	58.960	-112.10	177.97	163

Note: GDP per capita and the size of research sector in logs. N – number of observations.

Appendix Table 1.7: Definition and sources of the variables

Variable	Definition	Source
Predatory journal articles	The proportion of articles in journals linked to Beall's lists by authors from the respective country in total articles from that country recorded in the Scopus database.	Scopus (2018a)
GDP per capita	Gross domestic product (GDP) converted to constant 2011 international dollars using purchasing power parity (PPP) rates.	World Bank (2018)
Size of the research sector	Counts of total articles by authors from the respective country recorded in the Scopus database.	Scopus (2018a)
Oil and natural gas	The difference between the value of crude oil and natural gas production at regional prices and total costs of production as % of GDP.	World Bank (2018)
English spoken	Dummy with the value 1 if more than 20 % of population speaks English.	Mayer and Zignago (2011)
French spoken	Dummy with the value 1 if more than 20 % of population speaks French.	Mayer and Zignago (2011)
Spanish spoken	Dummy with the value 1 if more than 20 % of population speaks Spanish.	Mayer and Zignago (2011)
Arabic spoken	Dummy with the value 1 if more than 20 % of population speaks Arabic.	Mayer and Zignago (2011)
Latitude	Latitude of country centroid measured in degrees from the equator, with positive values going north and negative values going south.	Gallup et al. (1999)
Longitude	Longitude of country centroid measured in degrees from the Prime Meridian with positive values going east and negative values going west.	Gallup et al. (1999)

Chapter 2: Researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization

Co-authored with Martin Srholec, Márcia R. Ferreira, Nicolas Robinson-Garcia and Rodrigo Costas

Published as Macháček, V., Srholec, M., Ferreira, M. R., Robinson-Garcia, N., & Costas, R. (2022). Researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization. Science and Public Policy, 49(1), 85-97.

Abstract

We propose institutional mobility indicators based on researchers' mobility flows in 22 major fields of science across 1,130 Leiden Ranking institutions from 64 countries. We base our indicators on data from the Dimensions database and Global Research Identifier Database. We use researchers' first and last affiliations to estimate the extent authors have moved across institutions as well as countries. For each institution, we quantify the shares of researchers with the same affiliation (insiders), those who came from another institution within the country (domestic outsiders), and those coming from a different country (foreign outsiders). Institutions in Central, Eastern, and Southern Europe have the highest share of insiders, whereas institutions in Northern America and Western and Northern Europe have a higher share of foreign outsiders. Foreign outsiders are most common in small and wealthy countries. No disciplinary differences are observed, as captured by the field classification scheme of Dimensions.

Introduction

There is an increasing need to account for how institutions internationalize (Hoekman, 2012; Robinson-Garcia et al., 2019). Evidence suggests that cross-national mobility flows are increasing (Sugimoto et al., 2017) as is the strengthening of global collaborative research networks (Wagner and Leydesdorff, 2005). Research institutions foster internationalization as they compete to attract and retain foreign talent in a global market (Hazelkorn, 2011; Seeber et al., 2016), with foreign born researchers accounting for 43% of life sciences postdoctoral research in Europe, and 56% of these coming from outside the European Union (Moguérou and Di Pietrogiacomo, 2008). As a result, there is a policy interest in monitoring and understanding the process of academic mobility and internationalization of universities (OECD 2008, European Commission 2012, Jacob and Meek 2013, Sugimoto et al. 2017).

The lack of global and harmonized datasets has been a persistent challenge in developing global indicators of mobility (Sugimoto et al., 2016; Welch et al., 2018). This is especially problematic at the institutional level, which besides functioning author disambiguation mechanism requires clear institutional identification (Donner et al., 2020). Recently, bibliometric databases have significantly improved the consistency of publication metadata, particularly author-affiliation linkages, allowing to track affiliations of individual researchers and opening up new opportunities for studying long-term mobility patterns at scale (Moed and Halevi 2014, Sugimoto et al. 2016). Previous work has been concerned exclusively with the quantification of movement of scholars across countries and disciplines (Aref et al., 2019; Robinson-Garcia et al. 2019). However, evidence on institutional mobility has been lacking.

In this paper we aim to help in closing this gap by delivering novel insights on the composition of the academic workforce of universities and national research systems worldwide. We base our results on the Dimensions database and the Global Research Identifier Database (GRID). We propose three indicators by which institutional workforces can be characterized based on researchers' first and most recent institution of affiliation. By comparing the institution to which researchers were affiliated in their first publications with their most recent one, we distinguish between:

i) *insiders*, defined as researchers who are currently affiliated to the same institution to which they were affiliated in their first publications;
ii) *domestic outsiders*, that is, those who were originally affiliated to a different institution within the same country from their current one;
iii) *foreign outsiders*, researchers who were originally affiliated to an institution located in a different country from their current one.

The paper is structured as follows. First, we review the existing literature on academic mobility and the related concept of inbreeding. Then we describe the mobility indicators with a description of how we built the dataset. Next, we present the empirical findings from an aggregate perspective and analyze the heterogeneity of universities with regards to researchers' institutional mobility between countries and across the world. We conclude by discussing implications of our results, limitations of this approach and future research lines derived from this study.

Literature review

Academic mobility is widely perceived as beneficial to the scientific enterprise (Wagner & Jonkers, 2017; European Commission/EACEA/Eurydice, 2015). It is considered important to promote the dissemination of knowledge, acquiring experience in different research environments, career advancement, and for creating opportunities for collaboration (e.g., Sugimoto et al., 2017; Stephan & Levin, 2001; Wagner & Jonkers, 2017). Mobile researchers serve as important bridges between countries (Meyer, 2001), they reinforce existing or create new ties with other national and foreign institutions, thus generating collaborative research networks that span the globe. When researchers move, they bring with them knowledge and ideas that differ from natives, which are essential for knowledge recombination and interactive learning, which in turn may lead to innovation (Ganguli, 2015; Stephan & Levin, 2001).

Another important advantage of this connectivity is that it enables the division and coordination of labor and knowledge (Bettencourt, 2014). This is important as the search for scientific discovery becomes more complex, it increasingly requires a greater diversity of skills. These connections thus become vital to updating tasks, where a combination of different expertise is essential to ensure the completion of increasingly complex research projects. However, this also may lead researchers to become narrow specialists (Robinson-Garcia et al., 2020) who are less able to recognize relationships between separate groups of phenomena. The increased specialization of researchers combined with the need for attracting new knowledge and skills, increases research institutions' pressure to attract global talent. Despite the importance of mobility in science, much of the global characteristics of human capital distribution in the context of science remains underexplored.

The notion of academic mobility is closely related to the concept of internationalization (Wagner & Jonkers, 2017). The process of internationalization has become an essential part of universities' strategic planning as it involves the implementation of programs that support the periodic movement of researchers, building ties with top universities, and improving visibility. The increasing orientation towards internationalization is in part due to strategic considerations in the context of a global competition for talents (Seeber et al., 2016). Internationalization strategies are particularly used by universities in developed countries (Lepori et al., 2015), which are increasingly dependent on the movement of researchers as a way of maintaining their attractiveness and international reputation.

When deciding on which institution to go next, researchers tend to not only value the quality of prospective institutions, but also the attractiveness of the countries where those institutions are located play a significant role (Lepori et al, 2015). As a result, high income countries are much more likely to attract foreign researchers than developing ones (Chinchilla-Rodríguez et al. 2018). Socio-economic imbalances between receiving and sending locations has raised concerns among policy-makers that researchers' preferences and mobility may lead to a further marginalization of peripheral countries in the scientific landscape (Scott, 2015). The concern of human capital flight in 'source' countries (Stephan & Levin, 2001) has sparked the 'brain drain' (the loss of high-skilled workers) and 'brain gain' (the gain of knowledge workers) debate and is therefore a critical concern for policy makers (OECD, 2010).

The lack of institutional internationalization - which translates into low workforce mobility in a given institution - is thought to be decisive in influencing academic productivity and improving researchers' performance (Franzoni et al., 2014; Horta et al., 2010; Horta, 2013) and citation impact (Ganguli, 2015; Sugimoto et al., 2017). A recent large-scale bibliometric study reported that mobile researchers (more than one affiliation to different countries) were more cited than non-mobile researchers (Sugimoto et al., 2017). A similar finding was reported by Ganguli (2015) who found that Russian scientists who have migrated to the US have been more cited by US scientists than those who have not migrated. In the US context, foreign-born faculty has also been found to publish more (Mamiseishvili & Rosser 2010) and publish more breakthroughs (Stephan & Levin, 2001) than their US-born counterparts. One explanation for lower levels of impact and productivity of immobile researchers is that this kind of faculty has less access to varied information and exchange dynamics which may explain their lower scientific productivity (Horta, 2013).

Other studies suggest that academic mobility may or may not have a negative effect on the researchers' careers. For example, Melin (2005) has shown that mobile scholars returning to Sweden after a postdoctoral period abroad do not necessarily benefit from their foreign experience. Cruz-Castro and Sanz-Menéndez (2010) have found that non-mobile careers are a strong predictor of the timing of rewards in the form of early permanent positions in Spain, while mobile careers in the US increases researchers' chances of obtaining tenure. This suggests that the relationship between mobility and scientific performance varies widely across national research systems (Cruz-Castro & Sanz-Menéndez, 2010).

A related stream of literature contrasts mobility with the practice of having PhD graduates employed by the university that also trained them. This is known as “academic inbreeding” (Yudkevich & Sivak, 2012; Caplow & McGee, 1958; Berelson, 1960) or “institutional inbreeding” (Horta, 2013). Like the lack of mobility and internationalization, inbreeding is seen as negative for both the institution and researchers and has also been an indication of poor institutional performance (Horta, 2013). For example, academic inbreeding has been associated with lower output (Horta et al., 2010), fewer articles published in peer-reviewed journals than non-inbred (Horta, 2013), and fewer foreign co-authors (Scellato et al., 2012), effectively slowing down the career development of scholars (Inanc & Tuncer, 2011). Low productivity levels of inbred scholars emphasize the need for mobility and calls for policies to curtail academic inbreeding (Horta, 2013).

The university’s reliance on insiders among its workforce producing scientific results, can be interpreted as the university tendency to academic inbreeding. Empirical evidence on publication activity of PhD students indicates that even though differences between disciplines matter, the first papers’ affiliation can be used as a proxy for Ph.D granting institution of the individual researcher. For example, Lee (2000), Lariviére (2012) and Waaijer et al. (2016) found that PhD students publish their first article before graduation quite frequently in natural and health sciences, though less so in social sciences and humanities. Nevertheless, PhD students may acknowledge affiliation to their *alma mater*, even if the dissertation research is published only after leaving the university due to publication delays in peer-review journals.

Macháček and Srholec (2020) directly explored this measurement issue in a random sample of 90 researchers affiliated to major Western and Central European universities derived from the Scopus citation database. They found out from publicly available sources, from which university the researcher graduated, and compared that with outcomes of the bibliometric approach outlined below. The conclusion on whether the researcher is currently based on her alma mater matched by 77% in biochemistry, genetics and molecular biology, 90% in physics and astronomy and 87% in social sciences. There was the same number of 7 false positives and false negatives, which tend to offset each other, thus the impact in aggregated data is even more limited.

Data and methodology

Recent developments in bibliometric databases have made it possible to study career trajectories and aggregated patterns at the level of institutions, cities, and countries (Vaccario et al., 2020; Sugimoto et al., 2017; Robinson-Garcia et al., 2019; Murray et al., 2020). These developments include the implementation of advanced approaches for affiliation harmonization such as the Leiden Ranking approach, or, more recently, the Global Research Identifier Database, which currently covers more than 98,000 research institutions worldwide.

The availability of these harmonized registries makes it possible to identify changes in the affiliations of scholars (Moed and Halevi 2014, Sugimoto et al., 2016). When authors publish a paper with a certain affiliation, they signal that they are associated with that institution. This information allows us to track researchers' trajectories across institutions.

We use the Dimensions database version from June 2019 that is available in the database system of Centre for Science and Technology Studies on Leiden University (CWTS), including data on publications, affiliations, researchers as well as disciplines (Herzog et al., 2020), which in total contains data for more than 100 million documents. Dimensions cover slightly more documents than Scopus, but many publications lack affiliation links (Guerrero-Bote et al. 2021). Thelwall (2018) reports similar coverage of Dimensions to Scopus for publications with a Digital Object Identifier (DOI), which are predominantly journal articles but also book chapters, conference proceedings and others.

The Dimensions database has its own author name disambiguation procedure. In general, author name disambiguation algorithms attempt to group citation records of the same author by finding some similarity among them or try to directly assign publications to the individual who wrote them (Caron and van Eck, 2014). The Dimensions author name disambiguation algorithm is a two-step procedure. First, it uses "affiliation data, co-authorship and citation patterns as well as subject area traits" (Hook et al., 2018, p. 8) to produce clusters of publications that belong to potential individuals. These clusters are then connected using Open Researcher and Contributor ID (ORCID) and DOIs. Each disambiguated author is then assigned a researcher ID by which is unique identifier, producing almost 20 million researchers. Researcher IDs have been successfully assigned to about 87% of publications-authors combinations (Hook et al. 2018).

However, author-disambiguation algorithms (including the ORCID) are prone to errors (Caron and van Eck 2014, Gurney et al. 2012, Levin et al. 2012), and there is a need to establish a balance between precision (i.e. *How many identified publications truly belong to the same researcher?*) and recall (*Are all researchers' publications correctly identified?*). The Dimensions algorithm favors precision over recall (Bode et al. 2018), which is particularly relevant for us, since mobility events can cause splitting researchers into multiple IDs, and the precision preference can lead to the under-estimation of mobility events. For example, the publications of a researcher are split in disconnected clusters across her affiliations because the algorithm may not be able to merge them together.

Dimensions combines publication metadata with its own database of harmonized research institutions – the Global Research Identifier Database (GRID). This allows us to easily track the movements of scholars from one institution to the next. Aside from the unique institution identifier, GRID provides detailed geographical information, such as coordinates and city, region, and country names and codes. Other attributes include the institution's year of establishment and its type, which is divided into eight categories, one of which identifies educational institutions and universities.

Dimensions also includes a field-classification scheme for publications based on Australian and New Zealand Standard Research Classification (Australian Bureau of Statistics, 2008), which is a three-level hierarchical system of categories. Because more detailed disaggregation leads to a limited number of observations in the sub-disciplines, we use the top-level of this classification (FoR-division) in this paper, which refers to 22 major disciplines across all fields of sciences.

We used *2018* as the reference year to identify researchers affiliated with specific institutions. Furthermore, we only include researchers with sufficiently long publication histories (i.e., more than 6 years since their first publication). The year of first publication is used as a proxy for measuring academic age (Nane et al. 2017), preventing researchers with very short publication histories from driving the results. All researchers publishing their first paper in 2012 or later are excluded from the analysis.[22]

[22] Excluding researchers with short publication histories, and hence limited availability of relevant data, also mitigates the problem that researchers could be mistakenly broken down into multiple researcher IDs by the disambiguation algorithm.

Consider researcher r is affiliated to institution *u* and publishing in discipline *d* in the reference year. Each researcher r can belong to multiple sets as researchers can be affiliated to multiple universities and publish in multiple disciplines. She has a set of publications $P \subset p_{p,t}$ where at least one $p_{p,t>6}$, in which p is the publication index and t is the index for the number of years since the first publication (t is calculated from the publication calendar year). To determine where the researcher r started her career we select a subset of her initial publications published in the first two calendar years of her publication history, where:

$$P^{start} = \left|p_{p,t} where\ t < 2\right|$$

Based on P^{start}, we derive indicators on whether researcher r started her publication career at the same university ($r^{start,u}$) and country ($r^{start,c}$):

- Started on the same university?

$r^{start,u} = 1$ if r is affiliated to u in any publication in P^{start}, otherwise $r^{start,u} = 0$.

- Started in the same country?

$r^{start,c} = 1$ if r is affiliated to any institution in country *c* in any publication in P^{start}, otherwise $r^{start,c} = 0$.

This allows us to split researchers into three mutually exclusive categories. For each university *u* and discipline *d* we report the share on the total number of researchers of:

1. *Insiders*: Researchers starting at the same university – $\{\mathrm{r}|r^{start,u} = 1\}$
2. *Domestic outsiders*: Researchers starting at another institution in the same country – $\{\mathrm{r}|r^{start,u} = 0$ and $r^{start,c} = 1\}$
3. *Foreign outsiders*: Researchers starting abroad – $\{\mathrm{r}|r^{start,c} = 0\}$

We would like to include in this study only universities with significant research activities. The GRID category of "educational institutions" is far too broad for this purpose. Therefore, we use the Leiden Ranking methodology to identify the most productive research universities worldwide (Waltman et al, 2012). We use the Leiden Ranking 2020 data (see van Eck, 2020), which includes a total of 1,176 universities, only 7 of which we were unable to match with the initial set of GRID identifiers. While most of these universities have been created before World

War II and only about a tenth of them in the mid-1970s or later, there are some universities in that were established in 1998 or later according to the GRID database. We removed those universities from our sample since younger, less established universities have had less time to employ their own graduates[23]. The final sample includes 1,130 universities.

It is important to note that academic inbreeding is distinct from mobility. Our analysis explores the composition of an institution's workforce based on their initial and most recent affiliation, regardless of affiliation changes within the study period. This means that researchers labeled as *insiders* may have been mobile between those two points in time, but at the moment of the analysis they were linked back to their original institution. Adding a separate category for "returnees" could be insightful, but also complicated to define, and we leave that as a topic to explore in future research. Not accounting for returnees, however, is likely to overestimate the share of insiders particularly in environments with highly mobile researchers that are well connected to the global job market, and thus reduce the differences presented below.

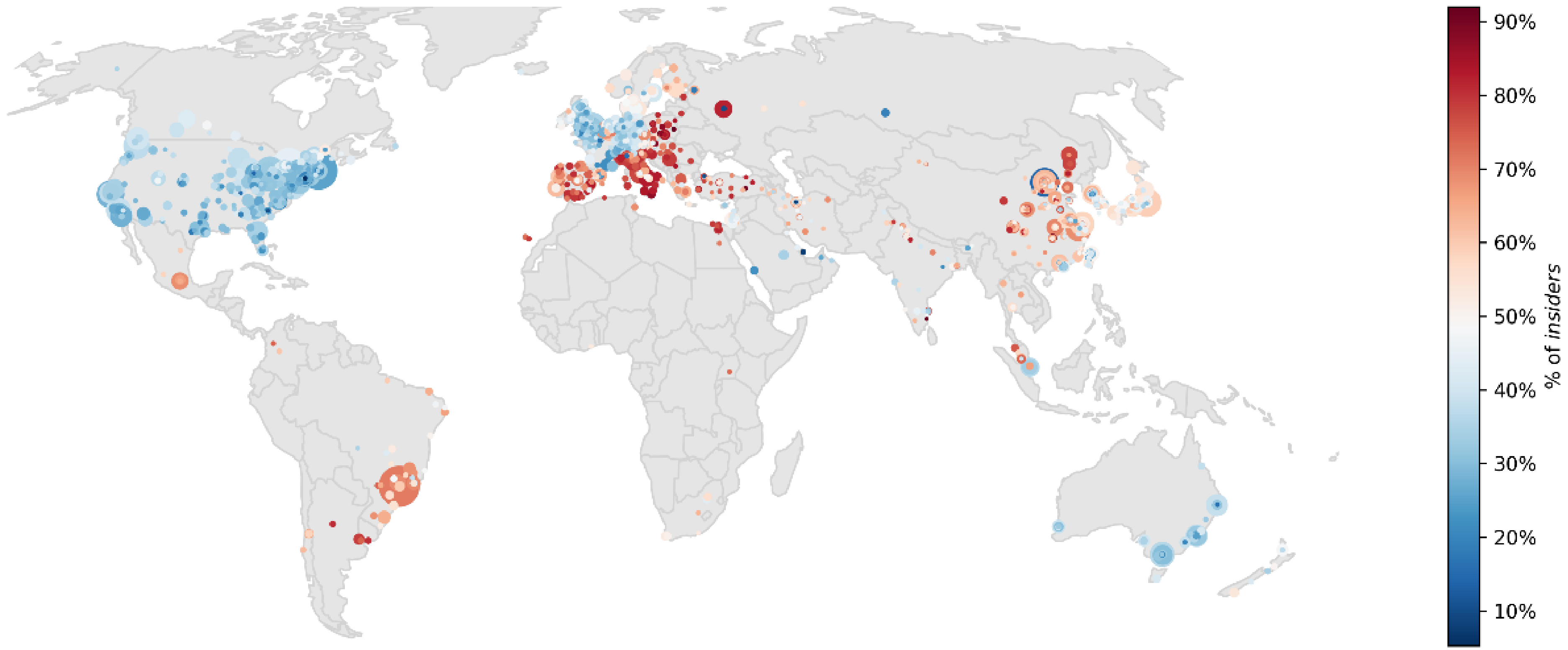
Also, we rely solely on author-affiliations linkages as reflected in researchers' publication record. That means that researchers will be labeled as *insider* regardless of the type of position they have at a given institution. Conversely, researchers are considered *foreign outsiders* for as many universities as they are linked as reflected by their publication record. Unfortunately, there is no easy way to identify all this diversity of affiliation linkages, unless one connects the bibliometric records with administrative data, which is not feasible at the global scale.

The share of foreign outsiders indicates the degree of internationalization of the university. Since researchers are more likely to circulate within the same area, because of institutional, cultural or personal obstacles for moving faraway, this distinction is probably more robust than between the insiders and domestic outsiders. If a researcher published early in her career solely with affiliation to a foreign institution, she was not likely to get research training in the same country, where she is based now. It is important to realize, however, that foreign outsiders may not necessarily be foreigners in terms of citizenship. It may well be that some of them only began their research career abroad and eventually returned to their homeland.

---

[23] A total 39 observations were dropped for this reason, of which 6 have the number of insiders equal to 0, and 5 exhibited less than 5% of insiders, which is below the minimum observed in the rest of the sample.

Figure 2.1: Share of insiders by university



Note: Only universities included in the Leiden Ranking and established before 1998 are included. Nodes denote universities. The size of nodes denotes the number of researchers identified. Colour reflects the share of insiders. Red indicates higher shares and blue lower shares of insiders.

Table 2.1: Descriptive overview of universities in the sample

	Mean	Median	St. dev.	Min	Max	Number of universities
Number of researchers	1,585	1,067	1,539	90	12,927	1,130
Institutional age in years	138	92	147	21	1,042	1,130
% of insiders	49.0	46.2	18.1	5.3	92.0	1,130
% of domestic outsiders	35.6	35.9	15.7	0.1	78.4	1,130
% of foreign outsiders	15.4	13.4	10.9	0.5	91.1	1,130

Note: Only Leiden Ranking universities that were established before 1998.

Results and discussion

Table 2.1 provides a descriptive overview of the dataset of the universities in the sample. The sample contains universities of different sizes, but even the smallest one has enough authors to reliably compute the indicators of our interest[24]

Overall, the share of insiders in total researchers differs markedly across the universities. On an average university nearly half of researchers hold the insider status, but this share ranges from less than a third in about one fifth of the universities to more than three quarters in a tenth of them; with some notable outliers at both ends of the spectrum. The tendency to employ initial outsiders also varies significantly with a clear predominance of locals over foreigners. Only about a tenth of the universities maintain more than a third of researchers who started publishing with affiliations abroad.

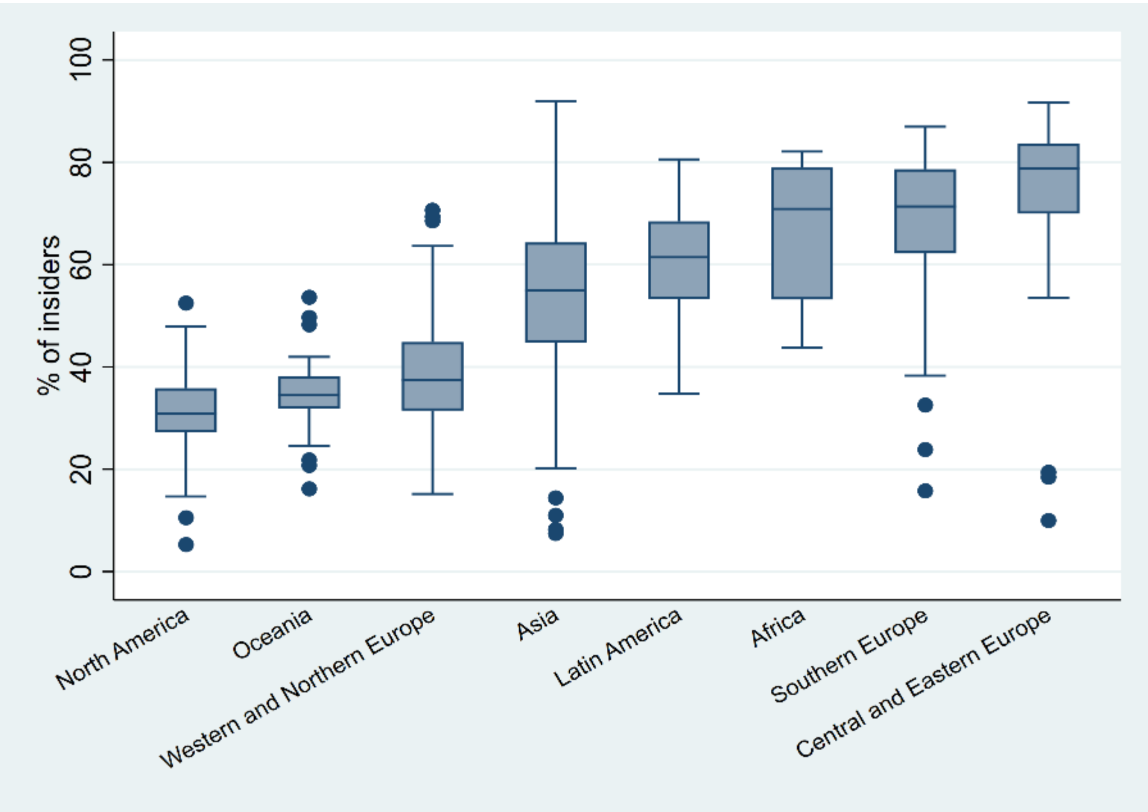
Figure 2.1 displays the share of insiders in universities worldwide. Each node denotes one university. The size of the node is proportional to the number of researchers in the respective university. The intensity of the colour reflects the share of insiders. Areas with the highest share of insiders are concentrated in the South and East regions, while low shares of insiders predominate in the North and West regions. Oceania, namely Australia and New Zealand, are the only exceptions. We also observe an overrepresentation of North American and European

[24] The number of authors with at least one publication affiliated with a university in 2018 (who published their first paper in 2012 at latest) could be used as a conservative estimate of its current number of researchers (i.e., its productive workforce). In fact, the actual headcount of researchers based at the university is likely to be significantly higher, as not everybody publishes every year

universities in contrast with other regions of the world such as Africa, Southeast Asia (except for China and Japan), and South America. Appendix figures 2.7 and 2.8 zoom the map into Europe and North America.

Figure 2.2 shows the distribution of universities based on the share of insiders in each geographic region. Europe, which is well represented in the sample, is further subdivided into three sub regions (Western and Northern Europe, Southern Europe and Central and Eastern Europe). The lowest proclivity to hire researchers originating from inside exhibit universities in the traditionally advanced countries of North America, Western and Northern Europe and Oceania. Within North America and Oceania, there is relatively little variation; half of the universities are boxed in a fairly narrow range, indicating that this is a systemic feature of how labour markets for researchers operate therein. Universities in Western and Northern Europe appear to be more diverse, with some high values, particularly in Scandinavian and Benelux countries.\

Figure 2.2: Distribution of the share of insiders in universities grouped by geographic region



Note: Only Leiden Ranking universities that were established before 1998. The box-plots depict median, boxes for interquartile range, whiskers with the length of 1.5 times that range and black dots for outliers falling outside of the whiskers

In contrast, a higher share of insiders is typical for universities in former soviet countries from Central and Eastern Europe, as well as in Southern Europe, where the tendency to employ insiders is roughly twice as high as in the Northern and Western parts of the continent. This shows that there are lingering differences within the European Research Area (ERA), even among the “old” member countries. In fact, Southern, Central and Eastern European universities have a propensity to employ insiders that is above of what is common in developing countries, including Africa, although evidence from the latter should be taken with a grain of salt due to a low number of observations from this continent.

For example, only one out of every four researchers currently affiliated to Harvard University, the University of Chicago, the University of Warwick and the Humboldt University of Berlin, but also the University of South Carolina, the Coventry University, or the University of Paderborn, started their career at the same institution. However, more than three out of every four currently affiliated researchers started publishing at Sapienza University of Rome, the University of Seville, the University of Warsaw, the University of Szeged, or the Moscow State University.

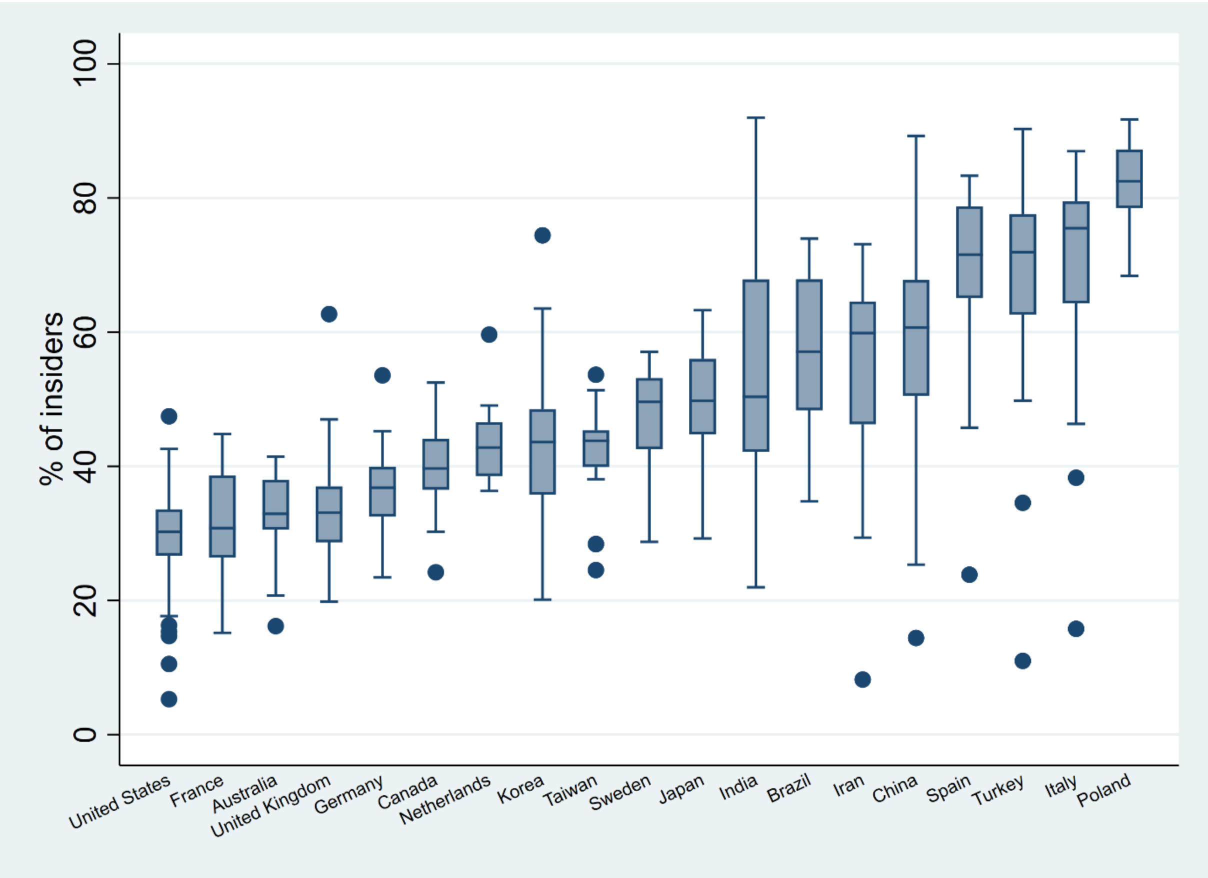
When we rank the continents by the median, Latin America and Asia fall between the two extremes. However, this masks vast differences within Asia, which is an amalgam of diverse countries ranging from advanced Japan to emerging China to developing India, but which proves difficult to divide by geography or other lines due to a low number of observations in the potential constituent parts and because the majority of them would be dominated by the largest countries. As illustrated below, the diversity within Asia is best understood by examining evidence from individual countries.

The outliers below the lower limit of the whiskers are predominantly elite local universities that seem not follow the suit in the higher shares of insiders that is common in their national environments. For example, this includes National Research University Higher School of Economics and ITMO University in Russia, Pompeu Fabra University in Spain, University of Cyprus and the University of Chinese Academy of Sciences. On the other end of the spectrum, with the share of insiders above the upper limit of the whiskers, the main outliers include Ghent University and University of Liège in Belgium or Åbo Akademi University in Finland.

Figure 2.3 complements these patterns by providing details on individual countries, for which results for at least 10 universities are available. Poland comes out clearly at the top of the list:

more than four out of five researchers are typically classified as insiders, and none of Polish universities goes below two-thirds of insiders. Next are Italy, Turkey and Spain with below the lower limit of the whiskers lower rates than Poland, but far more variability of the values. The largest variability is detected in China and India; both big countries with emerging and diverse university systems, where the share of insiders ranges from the highest figures in the world to the minimums observed in advanced countries. United States, France, Australia, the United Kingdom and Germany appear at the bottom of the list, which confirms that there is a strong developmental dimension in this ranking.

Figure 2.3: Distribution of the share of insiders in universities by country



Note: Only Leiden Ranking universities that were established before 1998. Only countries with results at least 10 universities. The box-plots depict median, boxes for interquartile range, whiskers with the length of 1.5 times that range and black dots for outliers falling outside of the whiskers

Table 2.2: Full results, sorted by insiders sorted by the share of insiders (% of total researchers, median)

	Insiders	Domestic outsiders	Foreign outsiders	Number of universities
Country groups:				
Central and Eastern Europe	78.8	14.7	5.8	59
Southern Europe	71.4	20.9	5.6	95
Africa	70.9	11.8	19.8	17
Latin America	61.4	27.5	7.8	44
Asia	54.9	32.1	9.2	448
Western and Northern Europe	37.4	40.9	22.6	204
Oceania	34.5	34.1	30.3	37
North America	30.9	52.9	16.1	226
Selected countries (10 or more observations):				
Poland	82.5	10.3	5.3	27
Italy	75.5	20.6	4.5	39
Turkey	71.9	20.4	6.1	30
Spain	71.6	21.4	6.7	41
China	60.7	30.2	8.9	201
Iran	59.9	33.4	8.5	36
Brazil	57.1	36.1	7.1	30
India	50.4	33.7	8.1	33
Japan	49.7	47.1	4.0	51
Sweden	49.6	23.2	25.6	12
Taiwan	43.8	44.1	13.4	21
Korea	43.6	36.2	19.3	42
Netherlands	42.8	34.2	21.8	13
Canada	39.7	32.2	25.1	30
Germany	36.8	45.2	19.1	53
United Kingdom	33.1	41.3	26.3	58
Australia	32.9	35.6	29.7	31
France	30.8	56.0	11.1	21
United States	30.2	53.6	15.3	196

Note: Only Leiden Ranking universities that were established before 1998.

Table 2.2 provides the median values and numbers of universities by geographic region and for countries featured in Figure 2.3, along with the share of domestic and foreign outsiders. The numbers of universities may appear disproportionate – even between countries of roughly similar sizes like China or India, and some large countries that are known to maintain extensive university systems are underrepresented in our dataset (e.g., Mexico and Russia).

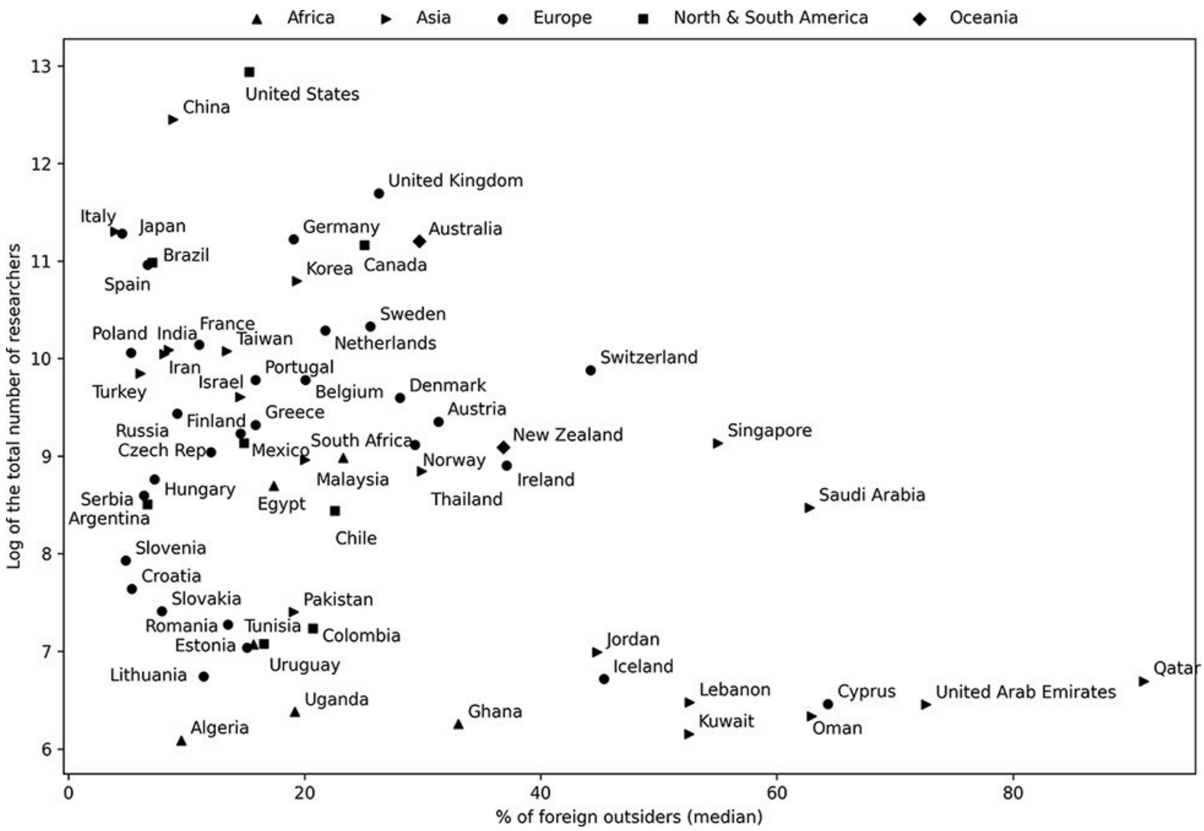
Shares of outsiders are negatively associated to the share of insiders, as they jointly add up to the total. More interesting is the proportion of domestic to foreign outsiders, and especially figures for the latter that provide insights on internationalization of the university system. Again, Southern, Central and Eastern European universities stand out with low shares of foreign outsiders as opposed to their Western and Eastern European counterparts. Likewise, developing countries reported in the table invariably show a relatively low share of foreign outsiders, while advanced countries generally exhibit the largest shares.

The share of foreign outsiders is negatively correlated to the size of national research systems. Thus, in large national research systems with many researchers, universities can draw from a large pool of domestic candidates, hence their chances to find a suitable candidate at home are naturally higher than in small research systems with a limited supply. For example, a university in Iceland is just for this reason likely to display a far higher share of foreign outsiders than otherwise similar university operating in the United States. Hence, one should compare in this regard universities from countries with research systems of a roughly similar size.

Figure 2.4 illustrates this point. On the vertical axis is the total number of researchers, regardless of the type of institution; used as a proxy for the size of the national research system. On the horizontal axis is the median share of foreign outsiders. All countries, which have at least one university in the sample, are displayed. The upper-right corner is empty and there is no country with a relatively large research system that displays more than a third of foreign outsiders, which confirms that there is a natural limit for large systems to attract foreigners over locals.

Size is clearly not all that matters. For example, Swiss universities are more internationalized, in terms of having foreign outsiders, as compared to other countries of similar size like Poland and Turkey, but also interestingly Israel, Portugal, Belgium and Denmark. Some high income countries, such as Japan, Italy and Spain also have a remarkable low share of foreign outsiders in their universities. The case of Japan is noteworthy (Table 2.1), since although it has overall modest share of insiders in its universities (49.7% on median), the largest share of their outsiders are domestic (47.1% on median), while they exhibit the lowest share of foreign outsider (4% on median).

Figure 2.4: Size of the domestic research system (log of the total number of researchers) and % of foreign outsider the share of foreign s (median)



Note: Only Leiden Ranking universities that were established before 1998.

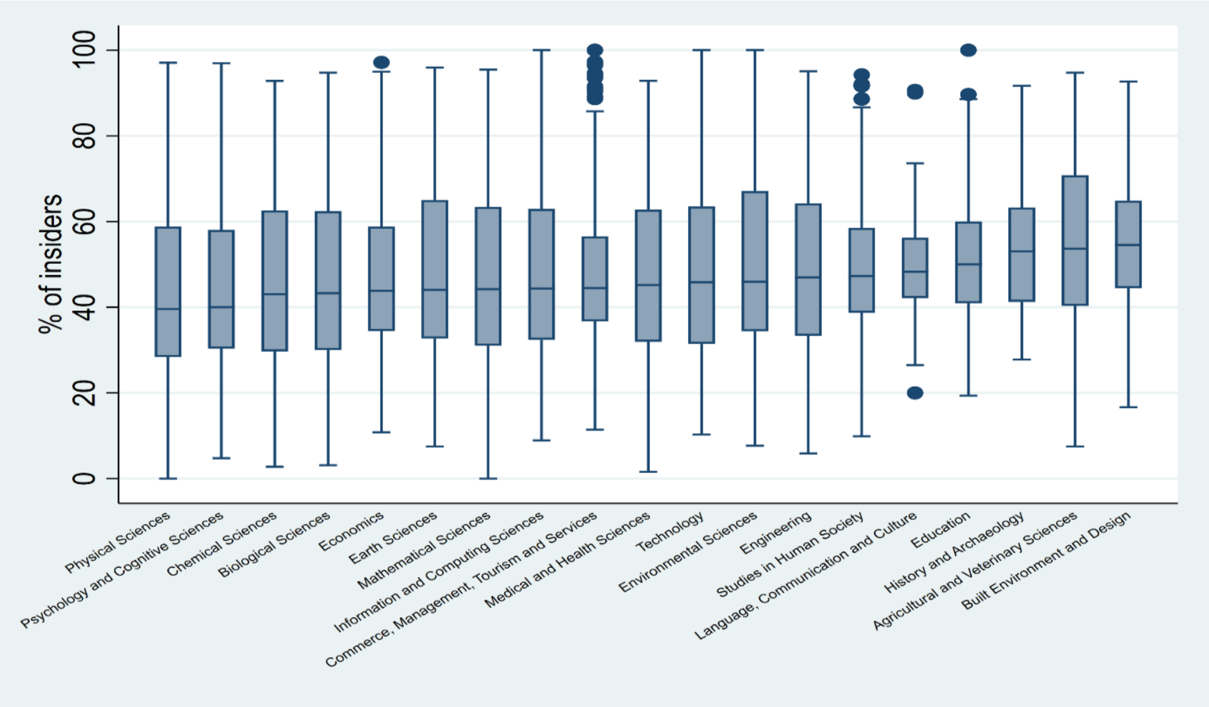
The largest differences are observed for countries with small research systems. Universities in countries like Slovenia, Croatia, Slovakia, Romania, Lithuania, Algeria, Pakistan and Uganda seem not to attract researchers who started their careers abroad, which is in contrast to Iceland, Cyprus, Lebanon or Jordan. Arabic oil-rich countries in the Persian Gulf, namely Kuwait, Oman, United Arab Emirates (UAE), Qatar and Saudi Arabia stand out with high shares of foreign outsiders.[25] This reflects their development strategy based on attracting foreign researchers (Schmoch et al., 2016), but this can also reflect the controversial strategy of universities in Saudi Arabia and elsewhere of offering secondary affiliations to highly cited

[25] In Figure 2.1 Saudi Arabia is notably one of the countries with more foreign outsiders. This country together with Qatar, UAE and Kuwait, have been described as “attracting countries” of researchers from abroad to their universities (El Ouahi, Robinson-Garcia, Costas, 2020).

researchers from abroad to boost their bibliometric profiles and ascend in rankings (Gingras, 2014). Arguably, this reiterates the limitation already mentioned above that we do not have any information on the parameters of contracts that underpin the affiliations of researchers in the published papers, although in practice these researchers contribute to the production (and visibility) of these universities.

Finally, we examine how the results differ by disciplines and thus to which extent these underlying differences could affect the patterns detected above. Using the Dimensions database, we can distinguish between 22 major disciplines across all fields of sciences, including social sciences and humanities. In order to present robust evidence, we narrow the sample only to universities with at least 30 authors in the respective discipline and present the results only for disciplines, for which such data is available in at least 30 universities. On these grounds, we eliminated from the analysis three disciplines (*Law and Legal Studies*; *Philosophy and Religious Studies*; and *Studies in Creative Arts and Writing*). The resulting dataset contains information on 10,408 pairs of university-disciplines of 1,129 universities across 19 disciplines.

Figure 2.5: Distribution of the shares of insiders by discipline



Note: Only Leiden Ranking universities that were established before 1998 with at least 30 authors in the respective discipline. Only disciplines with results for at least 30 universities are included. The box-plots depict median, boxes for interquartile range, whiskers with the length of 1.5 times that range and black dots for outliers falling outside of the whiskers

Figure 2.5 displays the boxplots for the share of insiders by disciplines. The main finding is that there is little variability across disciplines but a large diversity within them. The central tendencies are limited to a narrow range, the interquartile ranges are highly overlapping, while the whiskers tend to reach from close to zero to almost hundred percent in most disciplines. Arguably, this picture is in contrast to the significant differences that we have observed between universities across the national research systems above.

Table 2.3 further supports the conclusion that the university (and country) is the dominant level of the analysis with the help of variance components analysis. More specifically, we derive the intra-class correlation coefficient from an intercept-as-outcome mixed-effects linear model fitted via restricted maximum likelihood (Stata 2009, pp. 302-354). This procedure estimates the share of total variance explained by individual levels of the analysis. Stratifying the sample by disciplines explains less than 4% of variability in the data, regardless of the indicator, while differences between universities account for more than 70% and dividing the sample by countries explains more than 60%, respectively.

Table 2.3: Variance components analysis by country and discipline (% of total variance explained by individual factors)

Grouping	Insiders	Domestic outsiders	Foreign outsiders	Number of groups
Discipline	2.1	3.5	3.9	19
University	82.2	81.8	73.1	1,129
Country	67.2	60.3	83.3	64

Note: Only Leiden Ranking universities that were established before 1998 with at least 30 authors in the respective discipline. Only disciplines with results for at least 30 universities. Number of observations, i.e. university-disciplines, is 10,408. Proportion of variance is calculated as intra-class correlation coefficient from an intercept-as-outcome mixed-effects linear model fitted via restricted maximum likelihood (Stata 2009, pp. 302-354).

From this follows that organizational routines of universities for hiring, evaluating and promoting of researchers, which to a large extent reflect “rules of the game” given by national institutional frameworks, make the main difference, while particulars of individual disciplines do not matter that much. This also suggests that the main conclusions should be robust to differences in publishing practices between disciplines, including the propensity of PhD students (and graduates) to publish with affiliation to their *alma mater* that is important for interpretation of the results, as discussed above. Hence, the university as a whole, not necessarily the more nuanced university-discipline, is confirmed to be the most relevant unit of analysis.

A language of potential applicants can be an important factor affecting mobility decision. Universities in the country with a significant presence of a language with a billion speakers worldwide might have a larger chance to attract researchers speaking that language than universities in small countries speaking with just its own language. The language represents a factor that is beyond the control of universities and the research system in general. Other such factors can be other variables related to cultural proximity affecting general willingness to move, income difference between a host and a target country or administrative barriers to migration.

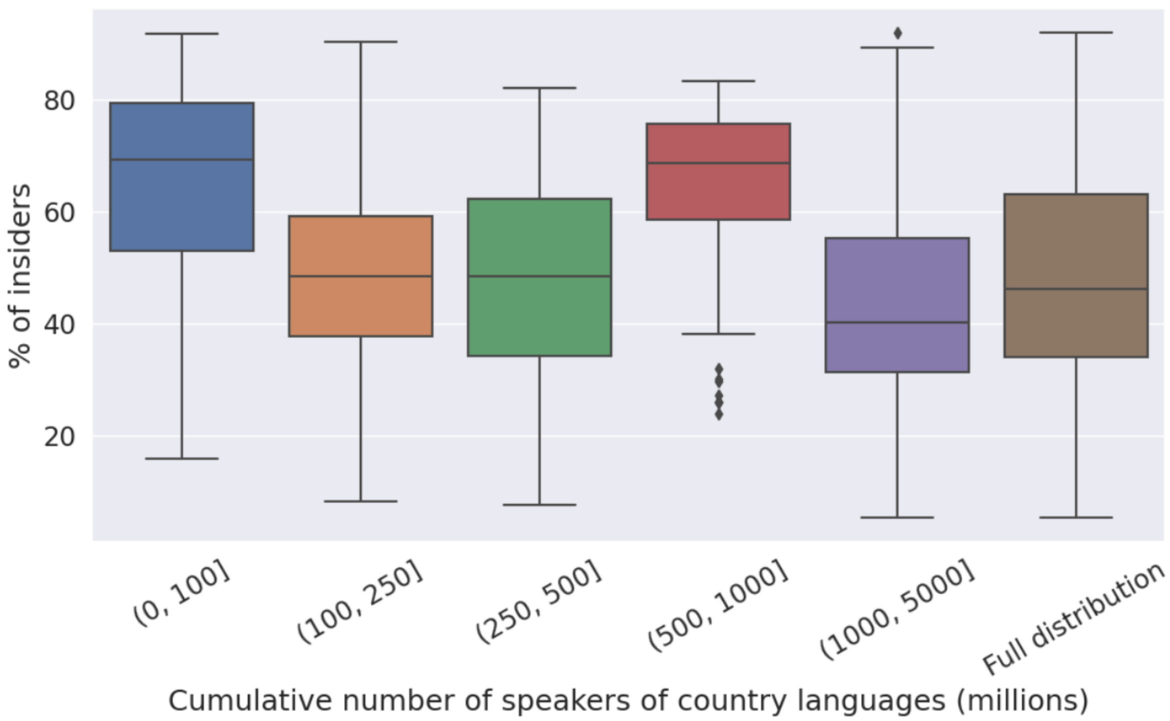
For each country in the dataset a list of languages spoken in the country was determined using a GeoDist database (Mayer and Zignago 2011). All languages including minority languages were included. Then for each language we found a number of its speakers on English version of Wikipedia[26]. When available, we used both native speakers (L1) and speakers as a second language (L2). Next, for each country a total number of speakers of their languages worldwide was determined by summing up the speakers of all languages in the country. For example in Singapur, there is significant presence of Chinese (1.12 bln. speakers), English (1.45 bln. speakers) and Malay (290 mln. speakers). Singapurian universities therefore have relatively easy access to the pool researchers from 2.86 billion speakers. On the contrary, there is only 1.1 millions speakers of Estonian language and no other languages is significantly present in Estonia according to GeoDist. Foreign researchers coming to Estonia might have to overcome a significant language barrier. In the final step countries were assigned to categories according to the number of speakers of the country languages.

58 % of all universities in the sample are in countries with *more than billion speakers*. All universities in the Anglo-Saxon world, but also other countries with strong ties to them – for example Korea, Israel, South Africa, Lebanon, Jordan or Pakistan – belong to this category.

[26] Although Wikipedia is not an ideal source for academic text, it serve its purpose in this specific case. Ethnologue database (see Eberhard et al. 2023) is beyond a paywall and no other comprehensive source was found. Manual extreaction of number of speakers from individual sources of would be time demanding with a very low value added in comparison with using English Wikipedia's figures.

Also speaking Chinese can help addressing an audience of more than a billion speakers – this is not only the case of China and Taiwan, but also of Malaysia and Singapur. Countries with *500 million to 1 billion speakers* consist of Spanish-speaking countries in Latin America and Spain and Algeria, Tunisia and Switzerland (6 % universities). *Between 250 million and 500 million speakers* is in French, Russian, Portuguese or Arabic speaking countries (7 % universities). German, Turkish, Persian and Japanese speaking countries can address a population *between 100 and 250 millions speakers* (17 % universities). The last category addressing *less than 100 million speakers* consists almost exclusively of small European nations – mainly countries of Central and Eastern Europe, but also Western countries such as Denmark, Sweden or Netherlands. The category is supplemented by Thailand (13 % universities).

Figure 2.6: Distribution of the shares of insiders by number of speakers of country languages



Note: Only Leiden Ranking universities that were established before 1998. The spoken languages in the country are from GeoDist database. The number of L1 and L2 speakers is from English Wikipedia (see Appendix Table 2.4 and Supplementary information 2.2). The box-plots depict median, boxes for interquartile range, whiskers with the length of 1.5 times that range and black dots for outliers falling outside of the whiskers

Figure 2.6 shows that only two categories deviate from the distribution of the whole sample – both countries with less than 100 million speakers and countries between 500 to 1 billion speakers seem to have somewhat higher share of insiders. The distribution in the rest of the categories is indistinguishable from the full distribution of the sample.

At the first sight, higher share in the first group could support a hypothesis that a small number of language speakers can hinder universities' ability to hire from outside. Countries in this category are relatively small European countries that do not speak with any of the world languages (supplemented by Thailand). However, deeper look reveals prevailing heterogeneity following the geographical patterns depicted on figures 2.1 and 2.2. Universities in Netherlands, Iceland, Sweden or Denmark tend to have a lower share of insiders. On the contrary, universities in countries of Central and Eastern Europe tend to have much higher share of insiders. Universities in Italy and Greece are somewhere in the middle with a considerable heterogeneity. The above average share of insiders in this category can be driven by factors not related to language.

The higher share of insiders in the latter category - countries with 500 million to 1 billion speakers – opposes the language hypothesis. Although it is the category with second highest number of speakers, almost all countries in the group - Spanish speaking countries as well as Arab and French speaking Tunisia and Algeria – tend to rely strongly on hiring insiders with the median share of over 60 %, sometimes even over 70 %. The only exception is Switzerland, where the share of insiders is among the lowest in the world with median below 30 %.

A preliminary analysis of language does not support the hypothesis that university hiring practices are associated by the number of speakers of language in the country. Although the analysis does not exclude the possibility that the relationship will become evident after controlling for interaction with other variables such as research evaluation culture. However, such analysis is currently unfeasible for the countries in the sample.

Conclusions

This paper demonstrates how bibliometric data can provide valuable insights into the institutional mobility of researchers. The empirical analysis reveals noticeable differences along the north versus south and west versus east geographical dimensions. The gulf between universities in Central, Eastern, and Southern Europe on the one hand and universities in North America, Western, and Northern Europe on the other is striking, pointing to systemic differences in how labour markets for researchers operate in the respective areas. Most developing countries fall between the two extremes, but there is significant variation within them. The findings also show that differences between universities and national research systems are most important, while disciplinary differences might be only marginally important.

The main findings on cross-country differences are broadly consistent with the growing body of empirical literature on geographical mobility of researchers, such as the MORE3 survey in Europe conducted by IDEA Consult, WIFO and Technopolis (2017) and the bibliometric analysis of mobility between European countries and the United States (Science Europe, 2013). Nevertheless, evidence on (a lack of) institutional mobility has been limited to qualitative analyses, surveys of particular contexts and/or case-studies of individual countries (Inanc and Tuncer 2011, Morichika and Shibayama 2015, Tavares, et al. 2015, Yudkevich et al. 2015, Horta and Yudkevich 2016 and Seeber et al. 2016), whereas broad comparative evidence based on harmonized data has been lacking. In this regard, the approach developed in this paper opens up new avenues for quantitative research as well as policy analyses on this topic.

The glaring differences between member countries of the European Union provide a sobering reminder that, even after two decades, the ERA agenda, which aims to create a single, borderless labour market for researchers with merit-based hiring, continental competition, and free movement of talent, still has a long way to go (European Commission 2018). It remains to be seen whether, for instance, expanding the EURAXESS network services or granting the so-called “HR Excellence in Research” badge to institutions that pledge to implement best practices will make a tangible difference, or whether far deeper structural reforms of university systems will be necessary to break the deadlock in the most autarkic countries.

Limited internationalization may reflect a lack of attractiveness of the national research system for relevant job candidates from abroad, which may be due to a relatively low national wage level in the first place. Nevertheless, in Central, Eastern and Southern European countries domestic labour markets for researchers do not seem to work well either, which is daunting especially in large countries with dozens of universities that produce a number of research job candidates who could circulate at least within the national system. Not only is the share of foreign but also domestic outsiders far lower in countries like Poland, Italy, and Spain than in France or Germany. These differences may point to excessively inward-looking hiring practices and general closeness to outsiders regardless of the extent to which these outsiders would be interested in getting the job. The larger presence of insiders in some countries may be explained by the combination of both endogenous (e.g. local research hiring practices, national career paths, etc.) and exogenous reasons (e.g. lack of attractiveness for foreign candidates, low visibility for foreign researchers, etc.).

As far as the developing countries are concerned, the results more than anything point to a large diversity, not only between them but most prominently to uneven development within some of the largest emerging national research systems. The main cases in point are China and India, but to a lesser extent also Iran and Brazil, where coexist universities with widely different profiles of researchers' institutional mobility. It will be interesting to observe which of these research systems converge towards the Northern and Western model of predominantly low academic inbreeding, or the Southern and Eastern model characterized by the opposite or whether they continue to be internally heterogeneous in a similar fashion as members of the ERA as the whole; and how this will change along the overall trajectory of their economic development.

Nevertheless, we cannot rule out the possibility that academic inbreeding could be beneficial in some circumstances, such as in newly forming fields of research carried in institutional and national contexts, when the stock of relevant outsiders may be limited, or in countries at low levels of development, for which attracting outside talent may not be feasible. Some research suggests that there can be upsides of academic inbreeding because local connections increase social capital of researchers, which benefit their careers (Yamanoi 2005 and Gorelova and Yudkevich 2015). As Woolcock and Narayan (2000) and other contributors to the literature on social capital point out; however, strong connections between people may well benefit them, but do not per se guarantee socially desirable outcomes. What becomes evident from this study is that scientometric approaches provide a fertile ground for quantitatively testing such hypotheses. The proof of concept presented in this study, clearly illustrates this literature would greatly benefit from taking this evidence more seriously.

The main take away from this paper for policymakers is perhaps that it is informative to use scientometric evidence for analyses of researchers' institutional mobility and that this line of enquiry ought to be deepened and extended along a number of lines. Admittedly, this study represents only a first glimpse into this direction. As already hinted above, future research should examine how these patterns evolve over time, and what is the impact of research policies on the mobility profiles of universities. Monitoring perspectives, which can be relatively easily implemented by repeating the same computational exercise for different periods, would be possible and advisable in the future, complementary to other approaches (e.g. MORE-4[27] and related surveys, etc.).

The researchers' decision whether to move or not combines a variety of factors from micro level (i.e. Do I want to move? Do I have an opportunity? Does it help my career?), the meso level (i.e. Are people surrounding me mobile? Are there organizational rules related to mobility?), and macro level (Are mobility policies implemented both in origin and host countries favouring the mobility decisions?). The factors on all three levels are mutually interrelated and it likely there are feedback loops reinforcing the macro level factors to micro and vice versa. More research is needed to understand how researchers exercise their mobility decisions in general, including a decision not to move at all. The decision stems not only from the micro-level of individual researchers, but is also affected by institutional policies and community practices on the meso level and national policies and cultural habits on the macro level. Such research may also better explain specific mobility flows[28].

Future research using this approach also needs to address some methodological limitations. The most pressing need is for examining in more detail the assumption that PhD students start to publish with affiliation to their *alma mater* that is crucial for interpretation of the insiders as inbreeders. We need to find out to which extent this assumption holds for researchers with different characteristics and from different contexts, for instance, with regards to multiple affiliations or part-time affiliations, to nail down the main sources of discrepancies and refine the computations accordingly. Another challenge that needs to be addressed is how to expand the coverage of the analysis beyond universities covered in the Leiden Ranking to bring forward even broader evidence, especially from developing countries, but without compromising coherence of the sample in the sense of mixing in the analysis academic institutions with excessively different missions.

Finally, it is important to remark that the approach presented in this paper follows a *backward-looking* perspective, studying the set of researchers currently publishing at a given university and looking backwards at where they started to publish. A different perspective would be a

[27] https://www.more-4.eu/

[28] A similar research proposal was submitted to this years' ERC Starting Grant 2023 Call by Aliakbar Akbaritabar.

forward-looking approach to studying mobility, in which universities would be characterized by the researchers who started to publish in them; and finding out whether this set of *alma mater* researchers continue at the same university afterwards or move elsewhere[29]. Although both approaches are complementary, it can be argued that both will capture different aspects of mobility, albeit with a different interpretation and implications, thus deserving future research about how they can be combined for a more complete picture about global academic inbreeding.

[29] Suppose a university that trained a lot of researchers in the past, half of them went abroad and the other half were employed by the same university, and that the university was not able to attract any other external researchers from other universities. In our *backward-looking* calculation this university would have a high share of insiders. Thus, our *backward-looking* approach wouldn't capture the fact that the university has sent abroad half of its trained researchers.

References

Australian Bureau of Statistics. (2008). Australian and New Zealand Standard Research Classification (ANZSRC). 13 Aug 2021. https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1297.02008?OpenDocument

Aref S, Zagheni E, West J. 2019. The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science In: Weber I, Darwish KM, Wagner C, Zagheni E, Nelson L, Aref S, Flöck F, editors. Social Informatics, Lecture Notes in Computer Science. Cham: Springer International Publishing. pp. 50–65. doi:10.1007/978-3-030-34971-4_4

Bettencourt, L. M. (2014). Impact of changing technology on the evolution of complex informational networks. Proceedings of the IEEE, 102(12), 1878-1891.

Berelson, B. (1960). *Graduate Education in the United States*. New York: McGraw-Hill.

Bode, C., Herzog, C., Hook, D., & Mcgrath, R. (2018). A Guide to the Dimensions Data Approach https://doi.org/10.6084/m9.figshare.5783094

Caplow, T., & McGee, R. J. (1958). *The academic marketplace*. Transaction Publishers. 1st edn. New York: Routledge .

Caron, E., & NJ van Eck. (2014). Large scale author name disambiguation using rule-based scoring and clustering. Available at: http://www.academia.edu/download/42806714/Research_quality_characteristics_of_publ20160218-10100-1wsawux.pdf#page=91

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). A global comparison of scientific mobility and collaboration according to national scientific capacities. Frontiers in research metrics and analytics, 3, 17.

Cruz-Castro, L., & Sanz-Menéndez, L. (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research policy*, 39(1), 27-38.

Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. Quantitative Science Studies, 1(1), 150-170.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2023. Ethnologue: Languages of the World. Twenty-sixth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

El-Ouahi, J., Robinson-García, N., & Costas, R. (2021). Analyzing scientific mobility and collaboration in the Middle East and North Africa. Quantitative Science Studies, 2(3), 1023-1047.

European Commision. (2012). Enhancing and focusing EU international cooperation in research and innovation: A strategic approach. 13 Aug 2021. http://ec.europa.eu/research/evaluations/index_en.cfm?pg=fp7

European Commission/EACEA/Eurydice. (2015). *The European Higher Education Area in 2015: Bologna Process implementation report*. Luxembourg: Publications Office of the European Union.

European Commission (2018) ERA progress report 2018. Luxembourg: Publications Office of the European Union. 13 Aug 2021. https://op.europa.eu/en/web/eu-law-and-publications/publication-detail/-/publication/5641328c-33f8-11e9-8d04-01aa75ed71a1

Franzoni, C., Scellato, G., & Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. *Economics Letters*, *122*(1), 89-93.

Gingras Y. (2014) 'How to boost your university up the rankings. University World News'. Jul 2014 https://www.universityworldnews.com/post.php?story=20140715142345754 .

Ganguli, I. (2015) Immigration and ideas: What did Russian scientists "bring" to the US? Unpublished manuscript, Stockholm School of Economics

Gorelova, O., Yudkevich, M. (2015) Academic Inbreeding: State of the Literature, in Yudkevich, M., Altbach, P. G., Rumbley. L. E. (eds.) Academic Inbreeding and Mobility in Higher Education. Palgrave Studies in Global Higher Education, Palgrave, London, pp. 17-44. http://dx.doi.org/10.1057/9781137461254_2

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative Analysis of the Bibliographic Data Sources Dimensions and Scopus: An Approach at the Country and Institutional Levels. Frontiers in Research Metrics and Analytics, 5, 19.

Gurney, T., Horlings, E., & van den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. Scientometrics, 91(2), 435–449. https://doi.org/10.1007/s11192-011-0589-1

Hazelkorn E. 2011. Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence. *Palgrave Macmillan*.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies, 1(1), 387-395.

Hoekman, J., 2012. Science in the Age of Globalization, The Geography of Research Collaboration and Its Effect on Scientific Publishing. Eindhoven University of Technology, Eindhoven.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. Frontiers in Research Metrics and Analytics, 3, 23. https://doi.org/10.3389/frma.2018.00023

Horta, H., Veloso, F. M., & Grediaga, R. (2010). Navel gazing: Academic inbreeding and scientific productivity. *Management Science*, 56(3), 414-429.

Horta, H. (2013) Deepening our understanding of academic inbreeding effects on research information exchange and scientific output: new insights for academic based research. Higher Education, 65, 487-510. http://dx.doi.org/10.1007/s10734-012-9559-7

Horta, H., Yudkevich, M. (2016) The role of academic inbreeding in developing higher education systems: Challenges and possible solutions. Technological Forecasting & Social Change, 113, 363-372. http://dx.doi.org/10.1016/j.techfore.2015.06.039

IDEA Consult, WIFO and Technopolis (2017) MORE3 study - Support data collection and analysis concerning mobility patterns and career paths or researchers. 13 Aug 2021. https://www.more3.eu/deliverables

Inanc, O., Tuncer, O. (2011) The effect of academic inbreeding on scientific effectiveness. Scientometrics, 88, 885–898. http://dx.doi.org/10.1007/s11192-011-0415-9

Jacob, M., & Meek, V. L. (2013). Scientific mobility and international research networks: Trends and policy tools for promoting research excellence and capacity building. Studies in Higher Education, 38(3), 331–344. https://doi.org/10.1080/03075079.2013.773789

Moguérou, P., & Di Pietrogiacomo, M. P. (2008). Stock, Career and Mobility of Researchers in the EU. JRC Scientific and Technical Reports.

Morichika, N. Shibayama, S. (2015) Impact of inbreeding on scientific productivity: A case study of a Japanese university department. Research Evaluation, 24, 146-157.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. In Scientometrics (Vol. 90, Issue 2, pp. 463–481). Springer. https://doi.org/10.1007/s11192-011-0495-6

Lee, W. M. (2000). Publication trends of doctoral students in three fields from 1965–1995. Journal of the American Society for Information Science, 51(2), 139–144. https://doi.org/10.1002/(SICI)1097-4571(2000)51:2<139::AID-ASI5>3.0.CO;2-1

Lepori, B., Seeber, M., & Bonaccorsi, A. (2015). Competition for talent. Country and organizational-level effects in the internationalization of European higher education institutions. *Research policy*, 44(3), 789-802.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. Journal of the American Society for Information Science and Technology, 63(5), 1030–1047. https://doi.org/10.1002/asi.22621

Macháček, V. and Srholec, M. (2020) Where do universities recruit researchers?, IDEA Study 1/2020. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. 13 Aug 2021. https://idea.cerge-ei.cz/files/RecruitingResearchers/

Mamiseishvili, K., & Rosser, V. J. (2010). International and citizen faculty in the United States: An examination of their productivity at research universities. *Research in Higher Education*, 51(1), 88.

Mayer, T. Zignago, S. (2011). Notes on CEPII's distances measures: The GeoDist database. CEP II, Working Paper No 2011 – 25, http://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877.

Melin, G. (2005). The dark side of mobility: negative experiences of doing a postdoc period abroad. *Research Evaluation*, *14*(3), 229-237.

Meyer, J.B. (2001) Network approach versus brain drain: lessons from the diaspora. International migration, 39(5), pp.91-110

Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. Scientometrics, 101(3), 1987–2001. https://doi.org/10.1007/s11192-014-1307-6

Murray, D., Yoon, J., Kojaku, S., Costas, R., Jung, W. S., Milojević, S., & Ahn, Y. Y. (2020). Unsupervised embedding of trajectories captures the latent structure of mobility. arXiv preprint arXiv:2012.02785.

Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. Journal of informetrics, 11(3), 713-729.OECD. (2008). The Global Competition for Talent: Mobility of the Highly Skilled. OECD. 13 Aug 2021. https://www.oecd.org/sti/inno/theglobalcompetitionfortalentmobilityofthehighlyskilled.htm

OECD. (2010). Measuring Innovation: A New Perspective. OECD Publishing. Paris: OECD Publishing

Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020) Meta-Research: Task specialization across research careers. eLife, 9, e60586

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. Journal of Informetrics, 13(1), 50–63. https://doi.org/10.1016/j.joi.2018.11.002

Science Europe (2013) Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility. Science Europe and Elsevier's SciVal Analytics. 13 Aug 2021. https://www.elsevier.com/research-intelligence/resource-library/comparative-benchmarking-of-european-and-us-research-collaboration-and-researcher-mobility

Scott, P. (2015). Dynamics of academic mobility: Hegemonic internationalisation or fluid globalisation. *European Review*, 23(S1), S55-S69.

Scellato, G., Franzoni, C., & Stephan, P. (2012). Mobile scientists and international networks (No. w18613). *National Bureau of Economic Research*.

Schmoch, U., Fardoun, H. M., & Mashat, A. S. (2016). Establishing a World-Class University in Saudi Arabia: intended and unintended effects. Scientometrics, 109(2), 1191-1207.Seeber, M., Debacker, N., & Vandevelde, K. (2016) Mobility and inbreeding in the heart of Europe. What factors predict academic career in Dutch-speaking Belgian universities. In Paper presented to the OECD blue sky forum on science and innovation indicators

Stephan, P.E., Levin, S.G. (2001) Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review 20, 59–79 https://doi.org/10.1023/A:1010682017950

Stata (2009) Longitudinal-Data/Panel-Data Reference Manual, release 11. College Station: Stata Press.

Sugimoto, C. R., Robinson-Garcia, N., & Costas, R. (2016). Towards a global scientific brain: Indicators of researcher mobility using co-affiliation data. http://arxiv.org/abs/1609.06499

Sugimoto, C. R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017) Scientists have most impact when they're free to move. Nature News, 550(7674), 29

Tavares, O., Cardoso, S., Carvalho, T., Sousa, S.B., Santiago, R. (2015). Academic inbreeding in the Portuguese academia. Higher Education. 69, 991–1006.

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? Journal of Informetrics, 12(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006

Van Eck, Nees Jan. (2020). CWTS Leiden Ranking 2020 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3997240

Vaccario, G., Verginer, L., & Schweitzer, F. (2020). The mobility network of scientists: Analyzing temporal correlations in scientific careers. Applied Network Science, 5(1), 1-14.

Waaijer, C. J. F., Macaluso, B., Sugimoto, C. R., & Larivière, V. (2016). Stability and Longevity in the Publication Careers of U.S. Doctorate Recipients. PLOS ONE, 11(4), e0154741. https://doi.org/10.1371/journal.pone.0154741

Wagner, C.S. and Jonkers, K., (2017) Open countries have strong science. Nature News, 550(7674), p.32.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International journal of Technology and Globalisation*, *1*(2), 185-208.

Welch EW, van Holm E, Jung H, Melkers J, Robinson-Garcia N, Mamiseishvili K, Pinheiro D. (2018). The Global Scientific Workforce (GTEC) Framework. STI 2018 Conference. Leiden (The Netherlands): Centre for Science and Technology Studies (CWTS). pp. 868–871.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., Wouters, P. (2012) The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.

Woolcock, M., and Narayan, D. (2000) Social Capital: Implications for Development Theory, Research, and Policy. World Bank Research Observer, 15(2), 225-250.

Yamanoi, A., 2005. The academic marketplace in Japan: Inbreeding, Grades and Organization at Research Universities. *Higher Education Forum*, 2, 93-114.

Yudkevich, M., & Sivak, E. (2012) University inbreeding: an impact on values, strategies and individual productivity of faculty members. Strategies and Individual Productivity of Faculty Members (January 31, 2012).

Yudkevich, M., Altbach, P. G. and Rumbley, L. E. (2015) Academic Inbreeding and Mobility in Higher Education. Global Perspectives. New York: Palgrave Macmillan.

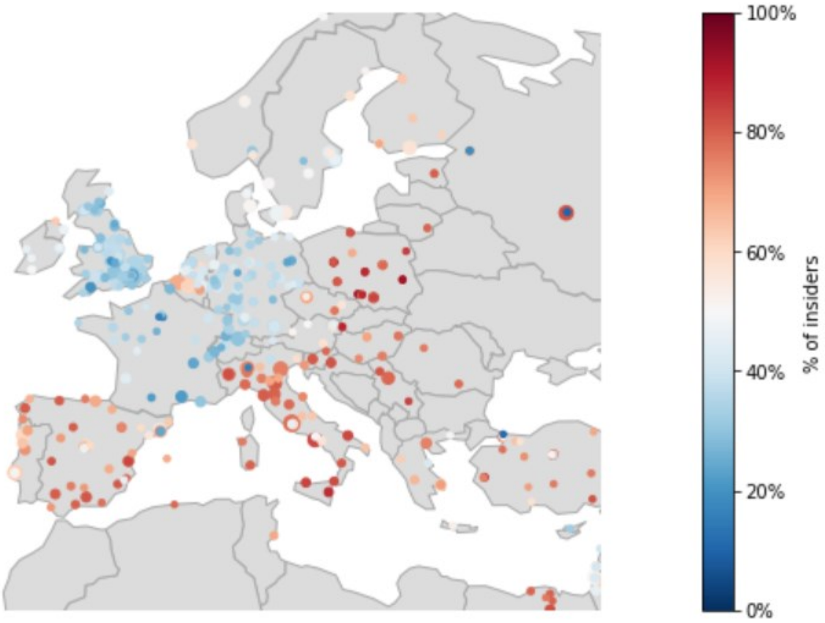
Appendix

Appendix Table 2.4 Number of speakers of country languages

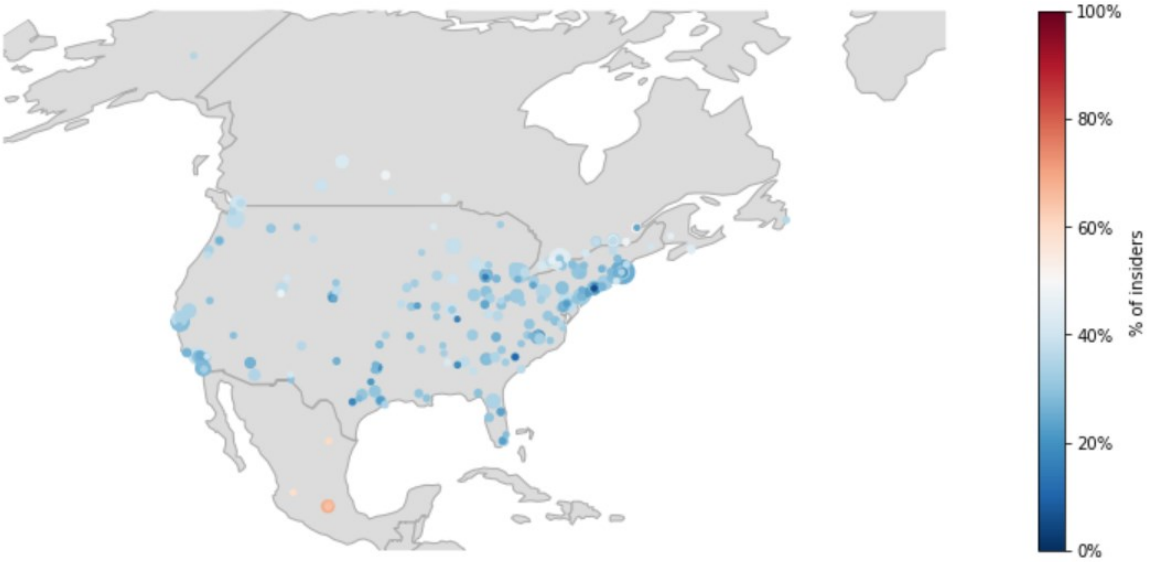
iso2	Speakers (mln.)	Languages	iso2	Speakers (mln.)	Languages
AE	360	Arabic	**JO**	1 812	English, Arabic
AR	595	Spanish	**JP**	128	Japanese
AT	180	German	**KR**	1 534	English, Korean
AU	1 452	English	**KW**	1 812	English, Arabic
BE	480	French, Dutch, German	**LB**	2 082	English, French, Arabic
BR	260	Portuguese	**LT**	3	Lithuanian
CA	1 722	English, French	**MX**	595	Spanish
CH	525	German, French, Italian, Alemannisch	**MY**	1 494	Tamoul, Chinese, Chinese Min Nan, Chinese Yue, Malay, Chinese Hakka
CL	595	Spanish	**NL**	30	Dutch
CN	1 118	Chinese Wu, Chinese, Chinese Yue, Chinese Mandarin	**NO**	5	Norwegian
CO	595	Spanish	**NZ**	1 452	English
CY	102	Greek, Turkish	**OM**	369	Balochi Southern, Arabic
CZ	11	Czech	**PK**	1 827	English, Punjabi, Urdu, Sindhi
DE	180	German	**PL**	45	Polish
DK	6	Danish	**PT**	260	Portuguese
DZ	634	Kabyle, French, Arabic	**QA**	490	Arabic, Persian
EE	1	Estonian	**RO**	46	Romanian, Hungarian
EG	1 812	English, Arabic	**RS**	27	Albanian, Serbo-Croatian
ES	607	Spanish, Galician, Catalan	**RU**	258	Russian
FI	19	Finnish, Swedish	**SA**	360	Arabic
FR	270	French	**SE**	13	Swedish, Skane
GB	1 452	English	**SG**	2 860	Chinese, Chinese Min Nan, Chinese Yue, Malay, English
GH	1 483	English, Akan, Éwé	**SI**	3	Slovenian
GR	14	Greek	**SK**	24	Slovak, Hungarian
HR	19	Serbo-Croatian	**TH**	90	Thai
HU	17	Hungarian	**TN**	630	French, Arabic

IE	1 454	English, Irish	**TR**	104	Kurmanji, Turkish
IL	2 349	Russian, English, French, Hebrew, Arabic	**TW**	1 118	Chinese Hakka, Chinese Min Nan, Chinese, Chinese Mandarin
IN	2 571	Marathi, Hindi, Bengali, English, Telugu	**UG**	1 488	English, Ganda
IR	154	Azeri, Persian	**US**	2 047	English, Spanish
IS	0	Icelandic	**UY**	595	Spanish
IT	78	Napoletano-Calabrese, Italian, Lombard	**ZA**	1 541	Shoto, Zoulou, English, Afrikaans, Xhosa

Appendix Figure 2.7 Share of insiders by university in Europe



Appendix Figure 2.8 Share of insiders by university in North America



Supplementary information 2.1

GitHub repository with source codes and public dataset:

https://github.com/vitekzkytek/ResearchersInstitutionalMobility

Supplementary information 2.2

Number of speakers for relevant languages (downloaded June 4th, 2023). Source Wikipedia pages included in the file

https://github.com/vitekzkytek/ResearchersInstitutionalMobility/data/languages_wiki.csv

Chapter 3: Globalization of Scientific Communication: Evidence from authors in academic journals by country of origin

Published as Macháček, V. (2022) Globalization of scientific communication: Evidence from authors in academic journals by country of origin, *Research Evaluation*, 2022; https://doi.org/10.1093/reseval/rvac033

Abstract

This study measures the tendency to publish in international scientific journals. For each of nearly 35 thousands Scopus-indexed journals, we derive seven globalization indicators based on the composition of authors by country of origin and other characteristics. These are subsequently scaled up to the level of 174 countries and 27 disciplines between 2005 and 2017. The results indicate that advanced countries maintain high globalization of scientific communication that is not varying across disciplines. Social sciences and health sciences are less globalized than physical and life sciences. Countries of the former Soviet bloc score far lower on the globalization measures, especially in social sciences or health sciences. Russia remains among the least globalized during the whole period, with no upward trend. Contrary, China has profoundly globalized its science system, gradually moving from the lowest globalization figures to the world average. The paper concludes with reflections on measurement issues and policy implications.

Introduction

Globalization of science is vital for addressing critical societal challenges (Wagner et al., 2015, 2017). Climate change, access to water or international fishing resources, control of infectious diseases, or social issues linked to development are just a few examples of topics where coordinated action from the international community based on scientific evidence is necessary. Globalization can bring more efficient allocation of research labor across the globe (Gui et al., 2019) and smooth scientific communication with efficient flow of ideas and feedback across borders.

Globalization of science is traditionally approached through *production.* Researchers engaging in international collaboration or academic mobility look for partnerships that help them produce papers more efficiently. We argue for taking a perspective of *scientific communication*, where globalization is derived from researchers' journal submissions' decisions. By sending a paper into a global journal, researchers send a signal that they want to communicate their paper with the whole world. On the contrary, submitting a paper into a journal with only local relevance indicates researchers' plan to communicate with a local audience.

Globalization of scientific communication (from now on just *globalization*) has been shown to be high in Advanced countries in Western Europe and North America (Gazni, 2015; Zitt & Bassecoulard, 1999; Zitt et al., 1998). Researchers in these countries by default publish in international journals. However, much more diverse picture will occur when looking beyond the most developed research systems in the world. In countries of the former Soviet bloc, there is a long history of local publishing which seems to survive to this day (Kirchik et al., 2012; Moed et al., 2018; Pajić, 2015). Local journals have an important role also in China (Zhang et al., 2021), Brazil (Brasil, 2021; Leta, 2012) or in Colombia (Chavarro et al., 2017).

The quantitative literature on the role of international journals in current research system is surprisingly scarce. Up to our knowledge, the only paper analyzing cross-country globalization patterns is Zitt & Bassecoulard (1999), published more than 20 years ago. Several papers analyzed internationalization of journals (Gazni, 2015; Gazni et al., 2016), but we are not aware of any cross-country comparison. At the same time, the global research landscape changed dramatically. It has grown both in size and interconnectedness (Royal Society, 2011; Science Europe, 2013) and collaboration distances increase (Waltman et al., 2011). International collaboration drives the growth of the research output (Adams, 2012, 2013). Developing countries invest heavily to improve their research infrastructure (Adams, 2013; Bornmann et al., 2015; Gazni et al., 2012; Wagner et al., 2015) and international visibility (Zhou & Glänzel, 2010).

This paper aims to bring fresh evidence on globalization of scientific communication. We use data on 34 964 journals indexed in Scopus and calculated globalization scores in 174 countries, 27 narrow and 4 broad disciplines between 2005 and 2017. In the text of this paper, we will only summarize the most important trends across countries, disciplines, and time. Nevertheless, the research community, policymakers, and other relevant stakeholders can benefit from the full detail of the rich dataset[30].

We build on Zitt & Bassecoulard (1998, 1999), who suggested journals' internationalization indicators based on bibliometric data and the procedure to derive globalization for the whole country and discipline. In this paper, we follow their procedure: 1) we assess the internationalization of all journals in the dataset, 2) in the scaling step, we analyze all published documents in the given country and discipline using the journal internationalization obtained in the first step.

The following section summarizes the literature on globalization of science, the approaches to its bibliometric measurements, and motivations to publish in international journals. Then, we describe our approach to the measurement of journal internationality and globalization of scientific communication. The data section describes a procedure to collect the data and representativeness issues. The results section and the discussion section contain a high-level summary of cross-country and cross-discipline comparisons. The last section concludes.

Globalization of science

Communication drives globalization. Scientific ideas, data, results of experiments and analyses, and feedback from peers abroad can be seamlessly distributed across the globe. Academic journals play a central role in this. They serve as critical communication platforms where ideas, results, and feedback meet. International journals facilitate cross-fertilization of ideas. As (Buela-Casal et al., 2006, p. 46) put it: "journals with wider national representation could increase the diversity of ideas and criticisms and be beneficial to the advancement of knowledge."

Scientific endeavor is increasingly globalized (Royal Society, 2011; Science Europe, 2013). The share of internationally co-authored papers on all papers more than doubled in twenty years (Wagner et al., 2015). The average distance between collaborating authors increased five times between 1980 and 2008 (Waltman et al., 2011). In advanced countries, the growth of

[30] Readers can take advantage of an interactive app globalizationofscience.com specifically designed for comparisons across countries and disciplines.

international collaboration is responsible for the actual growth of the research output (Adams, 2013).

Yet, the dynamics of globalization of science varies across the globe (Adams, 2012, 2013; Frenken et al., 2009; Gazni, 2015; Gazni et al., 2012; Gui et al., 2019; Robinson-Garcia et al., 2019; Royal Socety, 2011; Science Europe, 2013; Wagner & Jonkers, 2017; Zitt & Bassecoulard, 1999). The internationalization of the research system snowballed in the 1980s and 1990s in Western Europe and North America (Zitt et al., 1998). Since then, international collaboration still grows, and the barrier presented by national borders decays (Hoekman et al., 2010). At the same time, developing countries build new scientific capacities (Adams, 2012, 2013; Royal Society, 2011). The prime example of this are BRICS countries (Bornmann et al., 2015; Zhou & Glänzel, 2010; Zhou & Leydesdorff, 2006).

Globalization patterns also differ across disciplines (Gazni, 2015; Gazni et al., 2012; Royal Society, 2011; Science Europe, 2013; Zitt & Bassecoulard, 1999). However, the cross-discipline comparison can turn out to be challenging through the production-based measures as disciplines differ by the tendency to collaborate in general. Gazni et al. (2012) calculated shares of both multi-national and multi-authored publications across disciplines. They found 49 % multi-national publications and 83 % multi-authored publications in space science. These figures are 10 % and 42 % in social sciences. To a significant extent, the disciplinary differences in international collaboration are driven by the general collaboration patterns and team size (Mattsson et al., 2008; Wagner et al., 2017).

When researchers from multiple countries engage in international collaboration, they explicitly agree to produce science together. Each party offers her services – equipment, data, writing skills, labor, knowledge etc. Eventually they start working together and publish a paper. In this sense, collaboration can be perceived as a form of business deal, where researchers can only participate, if they have something to offer. Arguably, globalization measures based on international collaboration indicate the country's competitive advantage in science (Wagner et al., 2001).

Submitting an article into a journal sends a different signal. Journals' characteristics are a crucial evaluation input. The submission points to researchers' motivations and the reward system she faces. And researchers react strategically to the incentives (de Rijcke et al., 2016; Franzoni et al., 2011). When researchers know, they are expected to publish internationally, they will find a way to comply. On the contrary if the incentives favor publishing in venues that are disconnected from the rest of the world knowledge flows, there is a good chance that researchers will adjust the publication patterns accordingly.[31]

The production-based approach to globalization considers only a portion of the research output. Senior and elite researchers are more likely to engage in international collaboration and mobility (Czaika & Orazbayev, 2018). The less international the research is, the smaller fraction of the output enters the analysis. Claiming, for instance, only 10 % of researchers in the country engage in international collaboration neglects the globalization efforts of the remaining 90 %. On the contrary, communication-based measures can consider the entire distribution of the research output by analyzing the journal internationalization of each publication. If one is interested in how globally "ordinary researchers" act, globalization of scientific communication can be helpful, especially in the countries where international collaboration is less common.

The literature on local and international journals' role in the current research landscape is surprisingly scarce. In general, local journals seem to be in retreat. Gazni et al. (2016) analyzed 1 398 WoS-indexed journals during 1991 – 2014. They document a strong trend of journal internationalization throughout the whole period and across disciplines. Zitt & Bassecoulard (1999) documented the growth of international journals at the expense of local journals in all disciplines and all analyzed countries between 1981 – 1997, with the single exception of Russia.

At the same time, formerly local journals gradually internationalize. Gazni (2015) analyzed the development of WoS-indexed journals published by "national" publishers, defined as those publishing journals with official address in a single country. Authors from other countries than journals' publishers were marked as foreign. The share of papers with foreign authors grew from 36 % in 1990 to 62 % in 2013. The increase in internationalization was present in all disciplines, although significant interdisciplinary differences persisted in 2011-13. Journals also varied across regions – journals from Western Europe, the Pacific region, and Central Africa

[31] Obviously, also the international collaboration decision is affected by the incentives. Similarly, it would be naïve to separate globalization of scientific communication from scientific capacities entirely. The above distinction serves primarily to suggest main driving forces behind two interrelated concepts. It is beyond the scope of this paper to distinguish between them.

had over 70 % of papers with foreign authors, while the figure was below 40 % in Latin America (32 %), Middle East (36 %), Asian countries (46 %).

By submitting a publication to a journal, the researcher attempts to address certain community (Chavarro et al., 2014, 2017). Researchers balance the expected costs with the perceived benefits of publication in given venue. By submitting her work to an international journal, the researcher sends a signal that she wants to present to a global audience. On the contrary, by submitting a publication to a journal whose authors are predominantly local, she deliberately chooses to present to another, more localized community.

Non-international journals can be useful in their own right. Chavarro et al. (2017) offer several explanations why people publish in journals not indexed by standard bibliometric databases. Such explanation can easily be applied also to local journal. They identified three communication functions: 1) training of junior researchers, 2) knowledge bridging between mainstream science and local communities, and 3) publishing topics not well covered by mainstream journals. Ma (2019) adds speed and language of publication as an essential factor.

On the other hand, publications in local journals are likely to be less visible than publications in internationalized journals. Kirchik et al. (2012) analyzed the impact of Russian publications published in Russian journals translated to English to Russian publications in non-Russian journals. The number of citations in the latter group was much higher. The publication visibility goes beyond a simple publication language. It is not sufficient to just publish somewhere and expect that "*quality will prevail on its own*"[32]. Arguably, if the researcher wants to maximize visibility, she should choose more international journal.

The low internationality may be determined by the research topic. In many topics in social sciences and humanities, the object of the study is embedded in the local environment – consider language studies, history or development studies. Examples outside the social sciences include Chinese medicine (Ma, 2019), rice research (Ciarli & Ràfols, 2019), cassava, palm oil, passion fruit (Chavarro et al., 2017) or polar research. The topic is an important factor explaining lower globalization in specific disciplines. Nevertheless, it is unlikely to be driving force of the cross-

[32] See the case of Chinese journals in ISI (He & Liu, 2009; S. Wang et al., 2007).

country differences within the whole research system or broadly defined disciplines, especially in the large countries with a broad portfolio of research activities.

This paper is about the systemic tendency to publish in national journals. It shows the most flagrant cases of countries where the globalization of scientific communication is symptomatically low, regardless of the discipline. Researchers are affected by the incentives provided by the local research systems (Franzoni et al., 2011), which are primarily formed on the national level (Wagner et al., 2015). The resistance of local subjects towards internationalization policies can hinder globalization efforts (Zitt & Bassecoulard, 2004). A vital channel of this resistance can be through national journals, especially in research evaluation systems that rely on "*quantitative evaluation methods that are implemented in a formulaic and rigid manner*" (Rafols et al., 2016, p. 1). Anecdotal evidence of this can often be seen in the countries of the former eastern bloc (Bekavac et al., 1994; Good et al., 2015; Moed et al., 2018). The analysis of globalization of social sciences in Eastern Europe concludes that "it seems that the policies of (some) EE countries are too formal in stimulating futile publication behavior, aimed primarily at quantity, rather than quality" (Pajić, 2015).

Internationality should never be confused with quality. Bad journals can easily be highly international. On the contrary, neither high quality implies high internationality. The low internationality can be determined politically, culturally, and historically. Accusing Russian nuclear physicists of low-quality science would require much better evidence than low globalization of its scientific communication. There is much space for argumentation, exploration, and interpretation. Nevertheless, if there is systemic publishing in journals that are disconnected from the rest of the world's knowledge flows, the relevant stakeholders should be asking: *Is this what we want*?

Measuring globalization of scientific communication

Too often, journal internationalization is a loosely, if at all, defined term (Buela-Casal et al., 2006). There is no recognized boundary between "international" and "non-international" journals. There are many options to determine internationality. Indicators can differ by input, as well as by operationalization. The specific definition matters as varying definitions can lead to different rankings by internationalization. Buela-Casal et al. (2006) suggest creating a composite index to solve the ambiguity issue. We prefer keeping the individual indicators as separate inputs for the subsequent analysis to show that measurement artifacts from individual indicators do not drive the main patterns.

Internationalization indicators can be split into four categories according to its input: 1) user-based indicators analyze the community and feedback of users - journals subscribers, readers, or citers; 2) management-based indicators derive internationalization from the structure of the editorial board and publisher characteristics, and 3) author-based indicators consider the country of origin of authors; and 4) content-based indicators use published content as an input, such as language of publication. Due to the availability of data, this work will mostly rely on author-based indicators.

An appealing strategy to assess the journals' internationalization is to compare the number of contributions from the journal's domicile to the total number of papers. However, the major flaw of this approach is that for journals, and especially international, it is unclear what country is the journals' domicile. Arguably, this should be the country where most important editorial decisions are made in. The editorial board composition can be used to derive journals' domicile. However, especially in highly international journals, such a country may not even exist as strategic decisions are made literally globally.

Gazni (2015) uses publishers' location from Web of Science (similar information is also available in Scopus (2018)) as a proxy for the journal's domicile. However, potential misalignment between the publishers' official headquarters and journals' domicile can cause troubles when applying to the whole journal spectrum. To the best of our knowledge, there is no guidance of how the publishers' country is reported to databases. Especially for large multi-national publishers, the publisher country is sometimes puzzling[33].

Instead of vaguely determining journals' domicile, Zitt & Bassecoulard (1998) suggested that journals' internationalization indicators should compare journals' country distribution and the distribution of the entire discipline. Such indicators can naturally account for differences in the country's research sector size. They used distance-based measures comparing the country structure of the journal with the country structure of the aggregate discipline. When the distance between the two is low, meaning that the country structure of the journal closely resembles that

[33] In Scopus (2018) there is 2,349 active journals published by Elsevier; out of which 51 % is assigned to Netherlands, where Elsevier has its headquarters, 22 % in the United Kingdom and 16 % in the USA.

of the whole discipline, then the internationality is high. The further apart these distributions are, the more the journal deviates from the aggregate distribution, the less international the journal is. In other words, in an ideally international journal, a likelihood of getting published is independent of the author's country of origin. The less this holds, the less international the journal is.

Another option is to use concentration measures. They are somewhere in between the distance-based measures and measures based on journals' domicile. Sometimes it is convenient not to consider the global dimension of journal internationality. Consider the *European Journal of Public Policy* as an example. It should be highly international, but simultaneously Chinese and American researchers do not need to be equally represented as EU countries. These measures can, but do not have to, account for the size of the research output.

The rest of this section will present seven internationality indicators assessing each journal's globalization for each year. We included four author-based indicators using a country distribution of authors and one indicator using publication language. One indicator employs publishers' domicile, and one indicator measures institutional diversity – an openness indicator based on affiliation data. Three different data sources enter the analysis – country distribution of authors, institutional distribution of authors, and languages of publications. The indicators are not perfect, but each is imperfect in a different way. When combined, they yield a robust picture of the development of globalization.

Euclidian and *Cosine distance* are distance-based indicators. The difference is that cosine distance is independent of the magnitude (the vector length) and only compares angles. The ratio between each pair of countries is more important than the country's percentage share on all articles in the journal. Euclidian distance directly compares percentage shares. *Gini-Simpson index* was suggested by Aman (2016) as a journal internationalization measure "without a bias against periphery." It is a simple concentration measure used in ecology. *Largest Contributors Surplus* combines concentration and distance-based indicators. *Institutional Diversity* captures journal openness rather than internationality per se, and it uses affiliation data. We have also included a simple share of documents written in English (*English Documents*). The *Local Authors* represent journals' domicile-based indicators.

Table 3.1: Globalization indicators

Indicator	$I_{j,d,y,i}$ *calculation*	*Data*	*Indicator type*	*Source**
		Description		
Euclidian distance	$\sqrt{\sum (x_{j,y} - m_d)^2}$	Country	Distance	ZB (1998)
		Euclidian distance of journal and discipline country distribution		
Cosine distance	$\frac{\Sigma(x_{j,y} m_d)}{\sqrt{\Sigma(x_{j,y}^2)\Sigma(m_d^2)}}$	Country	Distance	ZB (1998)
		Cosine distance of journal and discipline country distribution		
GiniSimpson Index	$1 - \sum \frac{N_{c,j,y}^2}{(\sum N_{c,j,y})^2}$	Country	Concentration	Aman (2016)
		Gini-Simpson diversity of journal country distribution		
Largest Contributors Surplus**	$\sum_{c=1}^{3} (x_{c,j,y} - m_{c,d})$	Country	Concentration /Distance	Own
		Surplus of three largest contributing countries over its share in discipline		
Institutional Diversity**	$\sum_{i=1}^{3} (N_{o,j,y}/T_{j,y})$	Institutional	Concentration	Own
		Share of three largest institutions on all documents		
English Documents	$\frac{N_{ENG,j,y}}{T_{j,y}}$	Language	Index	BC et al. (2006)
		Share of English-written documents		
Local Authors	$\frac{N_{LOCAL,j,y}}{T_{j,y}}$	Country	Index	ZB (1998)
		Share of documents from a journal's domicile		

** ZB is Zitt and Bassecoulard; BC is Buela-Casal*

*** the underlying data for these indicators are sorted by descending order. The computation algorithm only considers the three most important*

Table 3.1 provides a detailed overview of the indicators. For each journal *j* in the dataset, a set of indicators *i* was calculated for each year *y*. The calculation is derived separately for each discipline *d*. $I_{j,d,y,i}$ denotes the journal internationalization. $N_{c,j,y}$, $N_{c,d,y}$ and $N_{o,j,y}$ are the number of documents with authors affiliated to the country *c* or organization *o,* in journal *j* or discipline

d, in year y. $N_{LOCAL,j,y}$ is the number of documents with authors from the same country as the publisher of journal j in the year y. $N_{ENG,j,y}$ is the number of English-written documents in the journal j in year y. $T_{j,y}$ denotes the total number of documents in the journal j in year y. Note that documents by authors from multiple countries are fully attributed to each country, i.e. $T_{j,y} \leq \sum_c N_{c,j,y}$. The vectors $x_{j,y}$ and m_d represent the country distribution of authors of the journal j and the discipline d, in which $x_{c,j,y} = \frac{N_{c,j,y}}{T_{j,y}}$ and $m_{c,d} = \frac{\sum_y N_{c,d,y}}{\sum_y \sum_c N_{c,d,y}}$.

Note that for distance-based indicators, the benchmark distribution m_d is calculated from all available periods. We benchmark the journals' distribution against the distribution of authors in the discipline during the whole analyzed period. Changes in journal distribution will therefore affect only $x_{j,y}$ and not m_d. The annual changes of the world trend will be fully attributed on the journals' side and not in the discipline aggregate.

Aggregation

In the second stage, the journal-level indicators were aggregated to the level of countries and disciplines. The resulting globalization score $G^S_{c,d,y,i}$ is a weighted average of individual journals scaled between 0 and 1, where 0 is the lowest globalization across all years, countries and disciplines within the indicator and 1 is the highest.

The globalization of science in country c, discipline d and year y expressed by an indicator i is calculated from the set of journals J assigned to discipline d as an average of individual journals globalization weighted by the share of documents flowing into the journal:

$$G_{c,d,y,i} = \sum_{j=1}^{J} \frac{N_{c,j,y}}{N_{c,d,y}} I_{j,d,y,i}$$

$\frac{N_{c,j,y}}{N_{c,d,y}}$ is the share of documents with authors from country c in journal j on all documents from the country c, discipline d in year y, $I_{j,d,y,i}$ is the globalization indicator i of journal j in the discipline d and year y.

Subsequently, the aggregated globalization index was standardized between 0 and 1 and converted to an ascending scale to simplify the interpretation of the results:

$$G^{S}_{c,d,y,i} = \frac{G_{c,d,y,i} - G^{min}_{i}}{G^{max}_{i} - G^{min}_{i}} \alpha_i$$

in which G_i^{min} and G_i^{max} is minimum and maximum value of the indicator i across all years, countries and disciplines and α_i equals -1 for the minimizing indicators (low values for high globalization) and 1 otherwise.

To increase results robustness and decrease volatility, the aggregation was only performed when the authors from the country published in at least 30 journals that published at least 30 documents in a respective year. This leads to gaps in results, particularly in the small disciplines and small countries.

Data

The analysis is based on the data from the Scopus citation database. Scopus indexes more journals than Web of Science (SCI-Expanded, SSCI, and A&HCI combined; see (Mongeon & Paul-Hus, 2016)). Hence, Scopus is more likely to contain the more local part of the scientific output in the country.

The data for all 34 964 journals indexed in the Scopus Source List (Scopus, 2018) were downloaded using Scopus API in August 2018.For each journal in each year between 2005 – 2017, we downloaded the country and institutional distribution of authors and the distribution of languages. Data were limited to articles, reviews, and conference papers. Scopus (2018) also contains the journals' publisher country as collected by Scopus.

The Scopus Search API was requested with the following query:

ISSN(AAAA-BBBB) AND DOCTYPE(AR OR RE OR CP) AND PUBYEAR = YYYY

in which AAAA-BBBB is the journal's ISSN and YYYY is the year. Rather than publication-level data, the aggregate distribution is collected. For each journal each year, we collect the number of articles affiliated to each country, language, and institution.

Scopus (2018) uses Scopus Journal Classification (Scopus, 2019) to assign journals to disciplines. We use *Major Subject Classification* of 27 disciplines (referred to as *narrow disciplines*), *Broad Subject Clusters* of 4 disciplines (life sciences, physical sciences, health sciences, and social sciences; referred to as *broad disciplines*), and an aggregate for all disciplines combined - *All*. The most granular level of Scopus classification – *Scopus Subject Areas* – was neglected due to concerns about representativeness and the risk of journals' false identification (Q. Wang & Waltman, 2016). The broad classification will be stressed in the rest of the paper, but the narrow results are also available in both the interactive application and the downloadable data.

Representativeness of data

Bibliometric databases such as Scopus are far from complete coverage of scientific communication. Wagner & Wong (2012) estimate that as much as 95 % of communication related to science (not necessarily peer-reviewed journals) is missing from the database. To be indexed in Scopus, journals must fulfill certain criteria (Scopus, 2022). Besides criteria linked to journals' management practice and academic trustworthiness, the journals are required to have geographically diverse composition of both authors and editors. They are also required to have English titles and abstracts. Minimum degree of internationality is expected from all the Scopus' content. Scientific communication platform with purely local validity should not be present and are not subject of this study.

Bibliometric databases probably represent a more significant portion of the science in academic journals in the countries of scientific core than those at the periphery (Chavarro et al., 2017; Ciarli & Ràfols, 2019; Mongeon & Paul-Hus, 2016); Zitt et al., 2003). With a reasonable assumption that the international journals are more likely to be indexed than non-international, the results can be interpreted as the upper bound of globalization.

The results are sensitive to Scopus journal-indexation decisions. For example, in 2009, Scopus reacted to criticism by increasing its coverage of social sciences and humanities journals by 39 % (Hicks & Wang, 2011). Longitudinal changes must be therefore interpreted with caution. Year-by-year jumps are not necessarily caused by fundamental changes in the researchers' behavior but are often driven by adding (or removing) journals in the database. Also, long-term changes may be driven by indexing a larger portion of existing journals.

The bibliometric databases cover disciplines unevenly as well. Mongeon & Paul-Hus (2016) report significant under-representation of social sciences and humanities in Scopus and WoS. Uneven representativeness is caused by the coverage of journals within the database. Also, for disciplines relying on other publication venues such as books, the results may be distorted. López Piñeiro & Hicks (2015) showed that significant share of research in Spanish sociology is published in journals not indexed in major bibliometric databases. A significant number of researchers, but even whole research topics might be overlooked. There is not much that can be done about this, but it should be kept in mind when interpreting the results. Again, assuming that the not-indexed part of the research is less globalized, we should interpret the results as the upper estimate of globalization.

Results

The computation algorithm yielded globalization scores $\mathrm{G}^{\mathrm{S}}_{\mathrm{c,d,y,i}}$ in three major dimensions - 174 countries (c), 31 disciplines (of which 27 *narrow* and 4 *broad* disciplines) (d) and the periods between 2005 - 2017 (y). The panel is unbalanced as the rule of publishing in at least 30 journals in a given country, discipline, and year is applied. In 2017, for example, the data were available for 171 countries in the total figures across disciplines*,* 125 – 155 in *broad* disciplines, and less than 100 countries in 21 out of 27 narrow disciplines. Naturally, the larger the research production in the country and the discipline, the more globalization scores are available. The data coverage grows in time, together with the growth of the research output and indexation of journals by Scopus. The globalization scores $\mathrm{G}^{\mathrm{S}}_{\mathrm{c,d,y,i}}$ (from now on just scores) are scaled between 0 and 1. 0 always represents the lowest globalization across all years, disciplines, and countries computed using the respective indicator, and 1 represents the highest.

The scores computed by different indicators are relatively strongly correlated (see Table 3.2). 18 out of 21 correlation coefficients exceed 0.5, and half of the coefficients are higher than 0.65. All of the correlations are statistically significant at 1% level. *Euclidian Distance*, *Cosine Distance, Gini-Simpson Index*, *Largest Contributors Surplus,* and *Institutional Diversity* are highly interrelated. Their correlations are at least 0.64, but also 0.8 is no exception. Note that these indicators intrinsically differ in their nature. They include both distance-based and concentration-based indicators, and *Institutional Diversity* even uses an entirely different dataset. The most representative indicator is *Euclidian Distance* with a correlation coefficient higher than 0.75 with all other indicators except *English Documents*. That is why, by default, we refer to it when not stated otherwise.

Only *English Documents* and *Local Authors* stand slightly aside from the rest of the pack. *English Documents* scores suffer from a highly skewed distribution of underlying internationalization of journals. 80% of journals publish documents almost exclusively in English (at least 95 % documents in English), and almost 90 % of all analyzed documents are in English. Only a tiny portion of data thus drives the results. *Local Authors* use vaguely defined journals' domicile, as already argued above. Note that "top five most prolific publishers account for more than 50% of all papers published in 2013" (Larivière et al., 2015). These publishers are likely to publish the most international journals where its domicile is the least meaningful.

Table 3.2: Correlation matrix of Globalization scores across indicators i

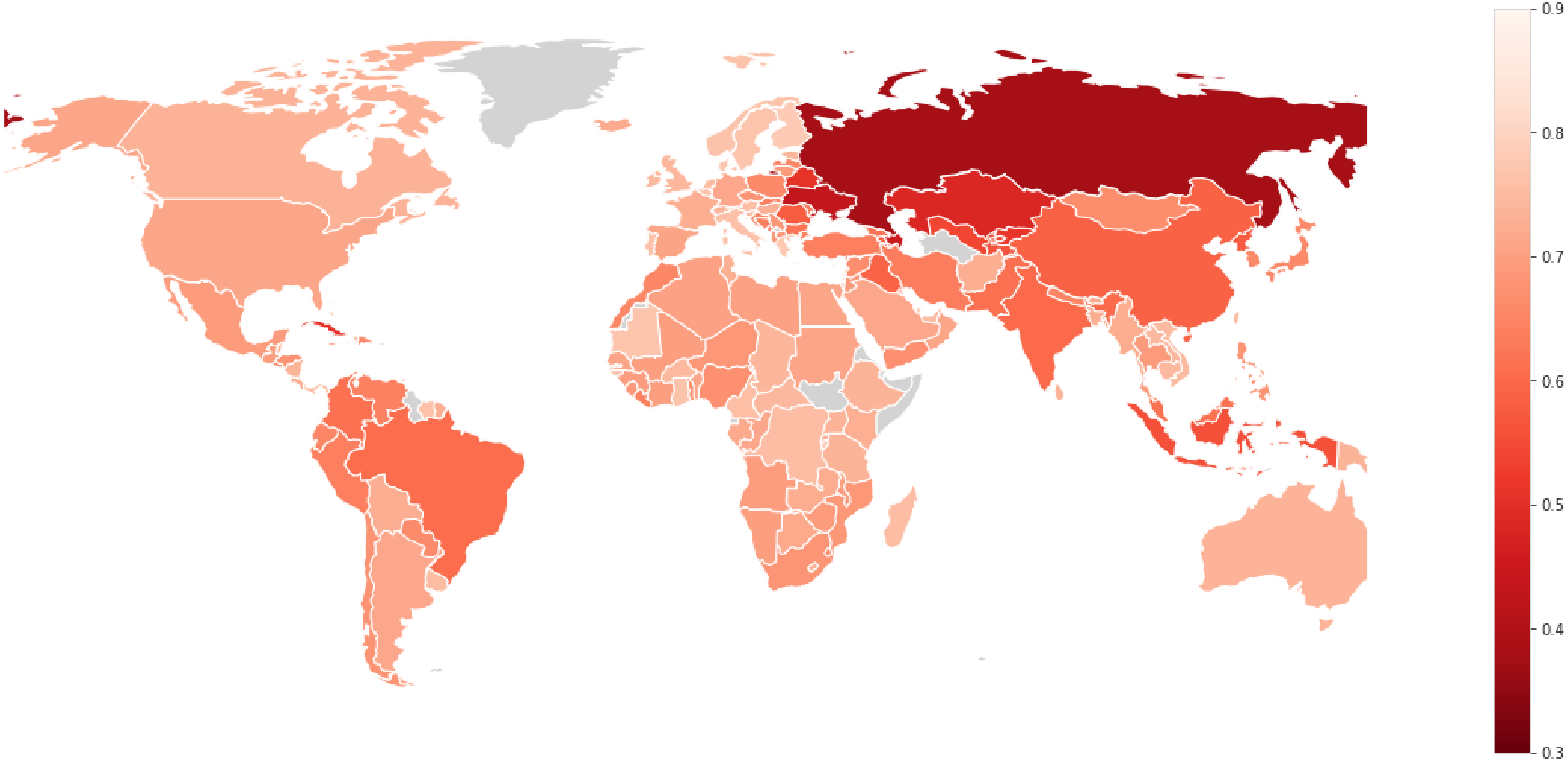
Indicator	*Cosine Distance*	*Gini-Simpson Index*	*Largest Contr. Surplus*	*Institutional Diversity*	*English Documents*	*Local Authors*
Euclidian Distance	**.83**	**.87**	**.93**	**.81**	**.61**	**.75**
Cosine Distance		.64	.75	.69	.47	.41
GiniSimpson Index			.72	.67	.64	.78
Largest Contributors Surplus				.79	.51	.67
Institutional Diversity					.43	.57
English Documents						.61

Source: Scopus; own calculation; Note: Pearson correlation coefficients of all available data for each indicator

Country and discipline differences

Figure 3.1 depicts total globalization scores in 2017 aggregated across disciplines as measured by Euclidian Distance on the map. The darker the color, the lower globalization. The results show a persistent differences between East and West. Globalization is high in Western, Northern and Southern Europe, North America, and Australia. On the contrary, the stronghold of low globalization is in the former Soviet Union – Russia, Kazakhstan, or Ukraine. Almost 30 years after the collapse of the Eastern bloc, the continuing isolation of their science systems is still very apparent.

Figure 3.1: Globalization scores on the map (2017, All disciplines, Euclidian distance)



Note: Grey denotes countries and regions with missing data. Source: Scopus, own calculation.

To a lesser extent, globalization seems to be lower in the BRIICS countries. In China, India, Indonesia, Brazil, and South Africa, but also, in other large countries in Asia – Turkey, Iran, Iraq, or Pakistan the globalization score $G^{S}_{c,All,2017,Euclid}$ is below the world mean of all country scores $G^{S}_{c,All,2017,Euclid}$. This also applies for Cosine distance, Gini-Simpson Index, Largest Contributors surplus and Institutional Diversity. Countries with emerging research infrastructure seem to be prone to lower globalization.

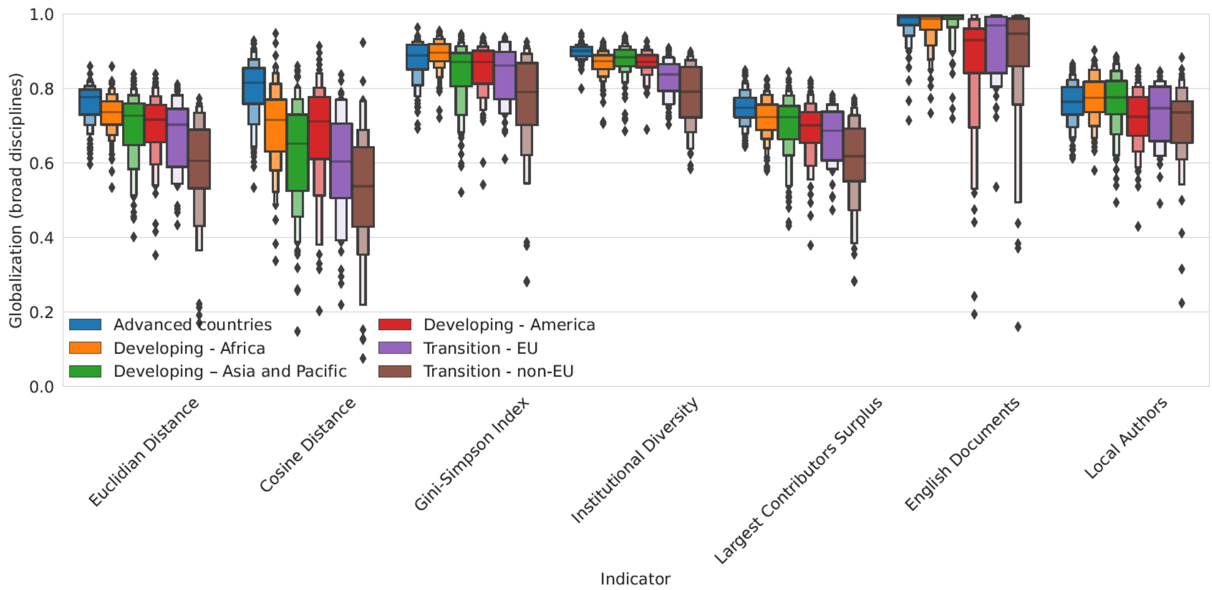
Interestingly, the least developed countries often have highly globalized science systems. In the Democratic Republic Congo, Mauritania or Cameroon, globalization is comparable with Great Britain, Netherlands, or Sweden. The same applies to globalization in Nicaragua, Costa Rica, Bolivia, or Surinam in Latin America or Myanmar, Afghanistan, and Vietnam in Asia. The production of science in these countries is often highly dependent on international collaboration. Moreover, the underlying data may be covering these science systems poorly.

The country groups proposed by IMF (2003) are used in more detailed examination by disciplines. The countries are divided into three categories: (1) *Advanced countries* cover the wealthiest countries in the world, mainly in Western Europe, North America, Eastern Asia, Australia, and Oceania. This group should capture the countries of the western core; (2) *Transition countries* consist mainly of the formerly socialistic countries in Central and Eastern Europe and Central Asia, including new EU member states; and (3) *Developing countries* – contain the rest of the world, including China. However, IMF (2003) classification is a rough brush with considerable within-group heterogeneity. Transition countries consist of the Czech Republic or Poland, right next to Kazakhstan, Russia, or Uzbekistan. Similarly, developing countries, for example, contain Saudi Arabia, China, Myanmar, and Bolivia or Democratic Republic Congo. Hence, we further split Transition countries by their EU membership and Developing countries by their continent[34]. The resulting country groups are: 1) *Advanced countries* (32 countries), 2) *Developing – Africa* (49 countries), 3) *Developing – America* (30 countries), 4) *Developing - Asia and Pacific* (35 countries), 5) *Transition – EU* (11 countries) and 6) *Transition - non-EU* (17 countries). The exact mapping of countries to country groups is available in Appendix Table 3.3.

[34] Only Malta was reassigned from Developing countries to Advanced countries

Figure 3.2 shows boxen-plots[35] across country groups in 2017. Each section contains distributions of scores from all broad disciplines computed with a respective indicator. Each "boxen" contains scores for different country groups.

Figure 3.2: Distribution of scores across indicators and country groups, 2017



Note: Each "boxen" contains scores for all broad disciplines within given country group in 2017. Source: own calculation, Scopus.

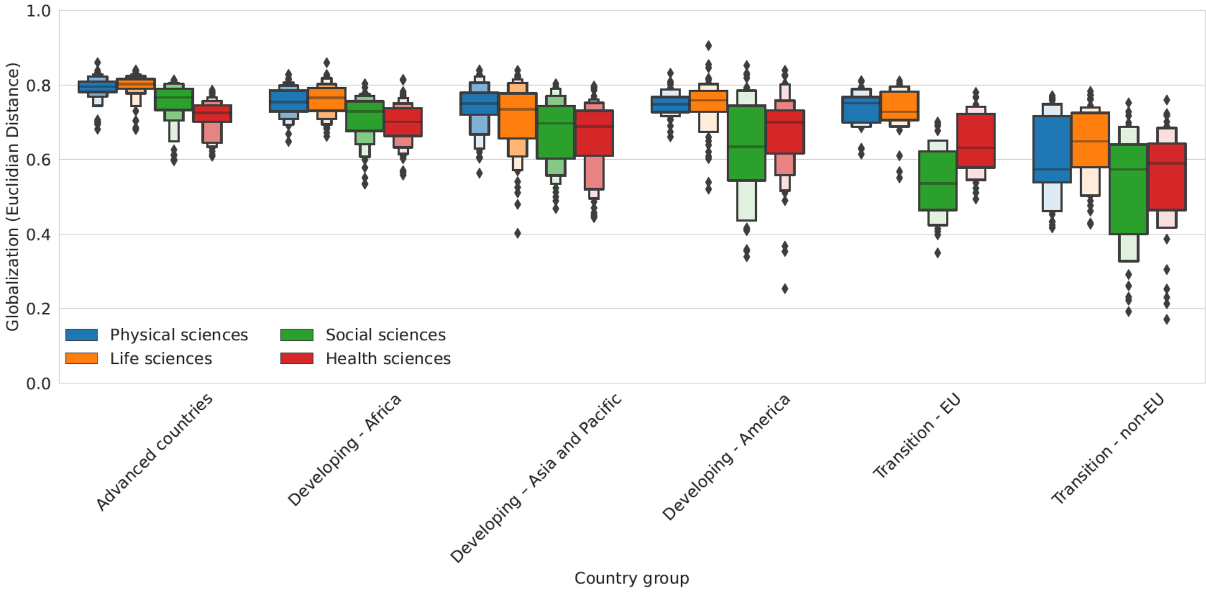
The figure confirms the pattern described above. The most globalized are *Advanced countries*. On the opposite side, the least globalized are *non-EU transition countries* – the group consisting mainly of the former USSR and Yugoslavia countries. Second-least globalized are *Transition countries from the EU* – the EU members from the former Eastern bloc followed by *Developing – Asia and Pacific* – fairly diverse group of countries such as *China*, *India*, *Indonesia, Turkey* or *Vietnam*.

Figure 3.3 shows disciplinary differences across country groups in 2015 - 2017. Each "boxen" represents the distribution of scores in one discipline and one country group. Three years were included to increase the number of observations. Having only one year does not qualitatively

[35] The line in the middle of each "boxen" represents a median. Each "*box*" of certain width represents a quantile of a certain distance from the median. The more distant quantile from the median, the thinner the box is. Each "boxen" depicts the entire distribution of globalization scores within a group of countries in given years.

change the results. Across country groups, *Life sciences* and *Physical sciences* are more globalized than *Social sciences* and *Health sciences*. What differ across country groups are the gaps, as well as the variance. *Life sciences* and *Physical sciences* are highly globalized in almost all country groups except non-EU transition countries. The differences are minor, if not negligible.

Figure 3.3: Distribution of scores across country groups and broad disciplines, 2015-2017



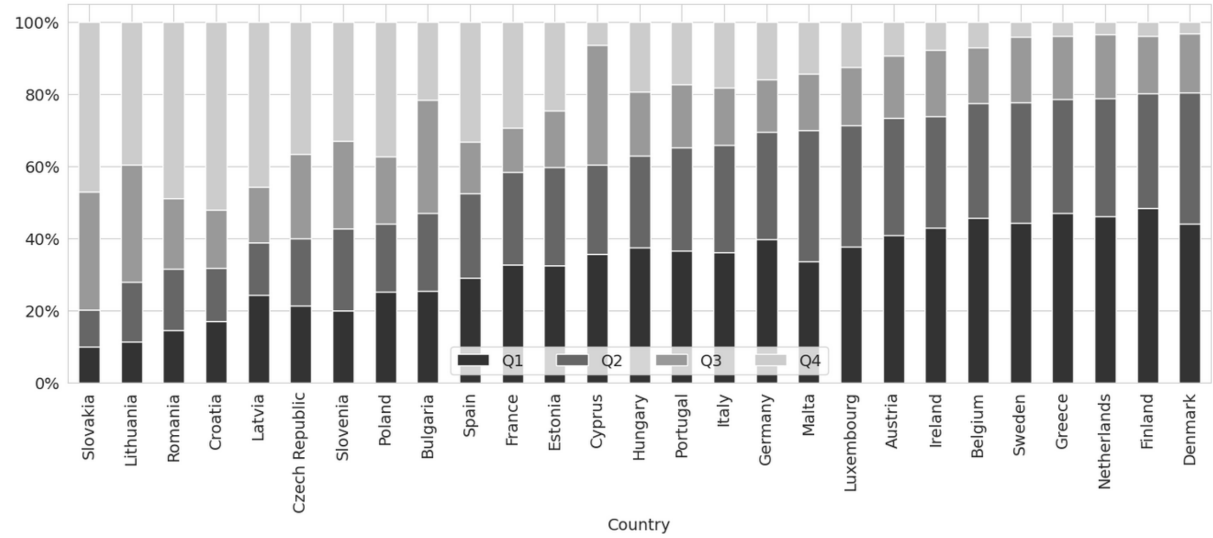
Note: Each "boxen" contains score for discipline and country group in 2015-2017 computed with Euclidian distance. Source: own calculation, Scopus.

Social sciences and *Health sciences* differ across country groups. In *Advanced countries,* they are almost equally globalized to *Life sciences* and *Physical sciences*. The gap between median of *Life* and *Physical* sciences and *Social* and *Health* sciences is wider in other country groups. At the same time, the variance of *Social* and *Health* sciences is larger. For example, in *Social sciences* in *Developing – America* in 2017, Brazil and Cuba are among the least globalized in the world (143th and 140th respectively out of 147 countries). The opposite holds for Panama (31st) or Jamaica (12th).

The large variance and low median apply to all disciplines in *Transition – non-EU* countries. A detailed look reveals that also this group contains two distinct groups – the former USSR countries and the former Yugoslavia. In former USSR countries, scientific communication is strongly isolated from the rest of the world. 8 out of 10 least globalized countries in discipline All in 2017 belongs to this country group. The Balkan countries' globalization tends to be much higher in all broad disciplines.

Social sciences in the EU

Figure 3.4: Distribution of documents into journals by globalization quartiles in Social sciences in the EU countries (Euclidian distance, 2017)



Note: Q1 represents 25 % of most international journals and Q4 25% least international. Source: own calculation, Scopus.

Social sciences stand out in *Transition - EU* countries relative to other disciplines in the same group (see Figure 3.3). Globalization in this context tends to be much lower than in other disciplines. At the same time, Social Science in the old EU member states (that belong to Advanced countries) is almost equally globalized to *Life* and *Physical sciences*. Hence, we provide a more detailed breakdown of the publication output into journals by internationalization in *Social sciences* in the EU.

First, all journals in *Social sciences* were split into quartiles by their Euclidian distance internationalization in 2017. In the second step, all documents from the given country and discipline were assigned to the journals' quartiles. Quartiles are marked Q1 – Q4, where Q1 is 25% of journals with the highest internationalization and Q4 with the smallest[36]. The darker the color in Figure 3.4, the higher the globalization.

[36] The closest journal to the median (journal right between Q2 and Q3) as measured by Euclidian distance in Social sciences in 2017 is International Spectator (ISSN 0393-2729), which published 35 documents. 12 had Italian affiliated authors, UK and USA based researchers contributed with 5 documents each. Other countries only had 1 or 2 documents. The closest journal to Q3/Q4 barrier -

In *Denmark*, *Finland*, *Netherlands*, *Greece*, *Sweden,* or *Belgium,* 80 % of documents were published in journals with above-median internationalization. This figure is only 20 % in *Slovakia*, 28 % in *Lithuania*, 32 % in *Romania*, 39 % in *Latvia,* and 40 % in the *Czech Republic*.

In some Western European countries, researchers rarely publish in the Q4 journals. In 10 countries this figure is below 10 %. In all new member states except Hungary, Estonia, and Cyprus, the Q4 journals accounted for more than 30 % of all documents published in 2017. Among old EU member countries, only France and Spain publish that often in Q4 journals.

Development over time

Changes in globalization over time must be interpreted with caution as the results combine changes in the researchers' publication behavior and the indexation decisions on the journals by Scopus.

In detail, we explore 12 large countries. We consider BRIICS countries – *Brazil*, *Russia*, *India*, *Indonesia,* and *South Africa*. These are supplemented by 6 largest countries (other than BRIICS) by a number of documents in 2017 – the *USA*, *United Kingdom*, *Germany*, *Japan*, *France,* and *Italy*, belong to the group of *Advanced countries*. Figure 3.5 depicts the development of globalization in these countries through the whole analyzed period.

The *Advanced countries* follow the trend of the whole country group. Globalization is high and invariant. The exceptions are the *Social* and *Health sciences* in France and, to a lesser extent, Germany. Countries speaking with what is often dubbed as "world" languages maintain an infrastructure of local journals in *Social* and *Health sciences*. They gradually globalize, but especially in France, many documents are still published locally (see Figure 3.4). Also, Japan, with somewhat lower globalization than its Advanced countries peers is slowly globalizing its research.

The case of *Indonesia* (and similarly of *Malaysia*) shows that the path towards higher globalization is not guaranteed. Globalization in these countries declines across disciplines. Further research is needed to ensure that this is not a data artifact of the Scopus' indexation decisions. However, it might be that hand in hand with building national research infrastructure, these countries gradually

---

Zeitschrift fur Slawistik (ISSN 0044-3506) - published 30 documents, out of which 15 had German-affiliated authors, and 5 were Russian. Again, other countries only had 1 or 2 documents.

build their ecosystem of local journals. Over 10 % of Indonesian documents from 2017 were published in the *Journal of Physics: Conference Series.* This journal published almost 30 % of documents with authors from Russia and 12 % of documents had Indonesia-affiliated authors. An additional 8 % of Indonesian research was published in *Advanced Science Letters*, a journal with almost 80 % of authors from Indonesia and Malaysia labeled as "potentially predatory" by J. Beall and whose coverage was discontinued by Scopus after 2017.

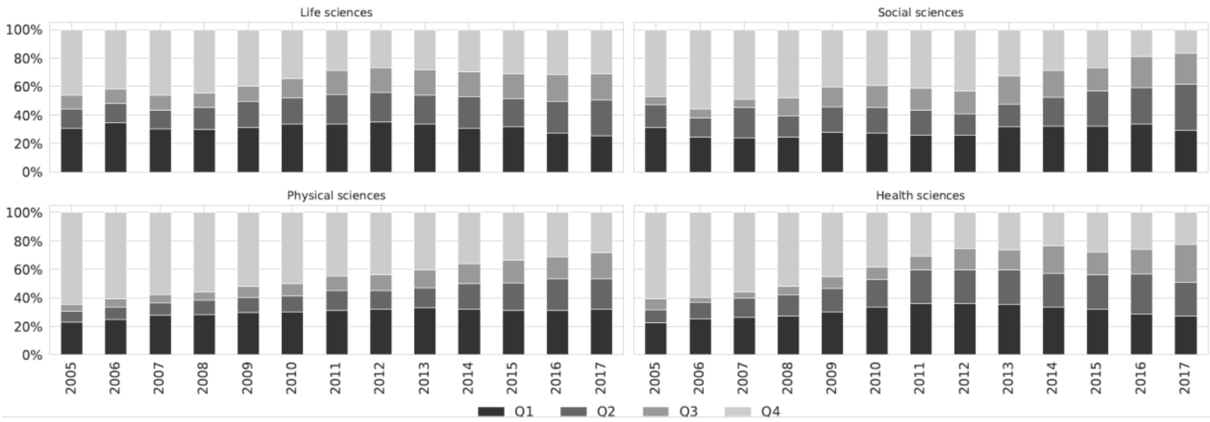
Figure 3.5: Globalization scores in BRIICS and other large countries over time, Euclidian distance

Source: own calculation, Scopus

This is in sharp contrast with the case of China. At the beginning of the analyzed period, in 2005, China was the least globalized country in the world according to the total figures and among the five least globalized in all broad disciplines (see also Moed, 2002; Ren and Rousseau, 2002). The rapid transformation of the Chinese system resulted in relatively fast growth of globalization. During the analyzed period, globalization grew fast across all broad disciplines.

The detailed breakdown of the Chinese research output (see Figure 3.6) reveals that the fast globalization in China can be attributed to the shift of publications from Q4 into the Q2 and Q3 journals. The share of publications in Q1 remains relatively stable in time. At least part of this development can be attributed to the relative growth of Chinese output. While in 2005, Chinese researchers contributed to approx. 8 % of all documents in the dataset, the same figure was almost 15 % in 2017. As a consequence, the internationalization scores of Chinese journals can grow not only due to different behavior of Chinese researchers, but also as a result of greater weight that China has as an international actor. That does not mean that Chinese researchers do not increase the presence in the international non-Chinese journals. Nevertheless, its pace is partly driven by the total growth of Chinese output.

Figure 3.6: Breakdown of research output across disciplines in China between 2005 – 2017 (Euclidian distance)



Source: own calculation, Scopus

Discussion

Globalization of scientific communication is generally high in the countries of the "core" – in Western Europe and North America. This finding is in line with Zitt et al. (1998), who announced an "*almost complete transition from national to the transnational mode of communication*" already at the end of the last century. Ponds (2009) hypothesized that international collaboration reached its upper limits in western countries like the Netherlands. The space for increasing internationalization is limited if any. Ponds (2009) predicted incorrectly that international collaboration in advanced countries is at its peak. But our data show that the prediction was pretty much correct from the perspective of scientific communication.

The opposite extreme are former USSR countries, where the standard publication model is centered on local journals. There is a long history of publishing in Russian journals (Kirchik et al., 2012). The low globalization is persistent, and it survives policies such as *Project 5-100,* which dedicated special funding to several Russian universities to "jump-off in the ladder" and rank at least five of them in the TOP 100 of respected rankings. Moed et al. (2018) describe an unexpected consequence of Project 5-100 – a fast inflow of Russian national journals into bibliometric databases after the program announcement. It can be challenging to make real change happen.

Doing science in isolated environment is a loss for everyone. Researchers lose visibility, feedback, citations, job opportunities abroad and potential future collaborators. But they are not the only losers. The rest of the world does not have a good access to findings generated within the isolated system. The journals might gain additional prestige if they opened up to the rest of the world. From the systemic perspective, intensive reliance on national journals effectively means that local researchers face lower competition from abroad. Unfortunately, it is likely that researcher lack not only competition, but also an inflow of ideas and feedback.

The high globalization of life and physical sciences is intuitive: "*There are no 'American' or 'Russian' electrons, atoms, or galaxies*" (Kirchik et al., 2012). The lower globalization in social sciences, where the object of study is embedded in the local environment, can also be quite natural. Let us take an example of economics. The economy works differently in each country as it is affected not only by the laws of supply and demand but also by institutional factors such as culture and regulation. One side of the discussion (let us label the argument as *contextual)* would stress the necessity to understand local specifics to the inner functioning of the economy. The opposite side – *universalistic* – would argue that the goal of science is to generalize rather than to describe. To isolate the effect of our interest, we need to extract it from its local embedding. We can only do so when we compare different contexts. If two countries differ in globalization within the single discipline, it can be explained in the perspective of *contextual vs. universalistic* debate. Where globalization is low, the researchers favor the *contextual* argument and vice versa researchers in highly globalized countries tend to practice their research in a more *universalist* manner.

Another explanation of the inter-country differences stresses the role of research assessment and the incentives from the local science systems (Franzoni et al., 2011). Country differences are stronger predictors of the resulting globalization than disciplines. The complicated question is what incentives matter and how they translate into higher or lower globalization. Do researchers facing a performance-based research evaluation system react by adjusting their submission decisions (Hicks, 2012)? Are there increased attempts to game the system via indexation of local journals (de Rijcke et al., 2016; Good et al., 2015; Moed et al., 2018)? It is not only about finance (Auranen & Nieminen, 2010; Quan et al., 2017). Successful research transformation requires a sensitive mix of ingredients. Of course, money is important (Franzoni et al., 2011; Quan et al., 2017), but other factors include teaching load, research evaluation requirements, quality of Ph.D programs, mobility or hiring policies (Kuzhabekova & Lee, 2018, 2020; Macháček et al., 2021).

Conclusions

The least globalized countries in the world should consider the outcomes of isolation. What is the purpose of having such a strong role of non-international journals? And what are the costs of isolation? If the costs are higher than benefits, how to reform the system efficiently? How to overcome the likely resistance of local subjects and how to support these who are willing to contribute to more open environment? These are all difficult questions that might take years or even decades to resolve successfully.

It cannot be stressed enough that the representativeness of underlying data limits our approach. Country's globalization may be linked to database coverage. The more research is missing from the Scopus database, the more likely it is that real globalization is in fact lower. We can hope that this shortcoming will abate with ongoing professionalization of science, digitalization, and improvement of bibliometric databases. But until then, the caution is warranted. The results are sufficiently robust to summarize major global trends in aggregate and broad disciplines. However, when digging into more detail – the situation in small disciplines and countries – it is essential to remain cautious and combine findings with the contextual information. In the meantime, the research community should also develop reliable and comparable indicators showing how much we do not know – share of major databases (at least Scopus and WoS) coverage on total country output.

An essential motivation behind this paper is informing policies. The rich dataset attached to this paper can be used to deliver policy-oriented studies. Each region, discipline, or country can be analyzed separately and combined with contextual information about the given research system. Hopefully, such studies will deliver comprehensive recommendations and targeted policies.

The literature will benefit from better understanding of how globalization of scientific communication relates to other measures of globalization of science. Various globalization measures – international collaboration, mobility, internationality of citations and of course internationality of journals - should be studied in the cross-country framework. What countries are strongly globalized from the communication perspective, but not that much from the perspective of international collaboration? And vice versa, in which countries researchers often engage in international collaboration, but publishing patterns are unexpectedly locally oriented? Globalization of science is a complex multi-dimensional phenomenon. Detailed understanding of its patterns and identifying the factors driving heterogeneity across indicators will help to better understand the link between research policies and its globalization outcomes.

Better understanding of the complex relationship between globalization and impact, especially in the context of individual career paths would certainly be beneficial. Chavarro et al. (2017) suggests that local journals can serve as “incubators” for junior researchers to learn the publication habits. The globalization dataset can be used to test whether these “learning platforms” are useful or junior researchers would gain (in terms of later impact) from being directly thrown to waters of international publishing.

To address internationalization policies, we need to understand within-country heterogeneity of researchers with respect to globalization. Taking the journals’ internationalization scores as an input, it is possible to study intergenerational patterns as well as whether local publishing differ across organizations and disciplines. Two independent clusters of nationally publishing and internationally publishing researchers would require different policies than if researchers rather mix – they publish some articles in international journals and others in local journals.

Another step further is to estimate globalization in higher disciplinary granularity. Using community detection algorithms (Traag et al., 2019), it is possible to identify much more granular disciplines than those of Scopus. This would allow to differentiate between topics that are predominantly linked to international journals and which topics typically end up in more local journals. Such research might help target policies to push internationalization only in topics, in which it is justified, without disrupting research that work best within the local environment.

On the contrary, globalization cannot become an uncontested dogma. There are serious concerns regarding its impact on developing countries. Naïve globalization policies that simply require publishing in journals indexed in important bibliometric databases may lead to the “westernization” of research in the country, as these databases predominantly index journals managed from western countries (Chavarro, Ràfols, & Tang, 2018; Larivière, Haustein, & Mongeon, 2015). Among the consequences of such westernization can be a blind adoption of research priorities dictated by editors of Western journals. Ciarli & Ràfols (2019) documented on the case of rice research that research priority settings do not always meet with the most pressing societal needs in the given the country. The internationalization pressures might be diverging attention of researchers away from the locally important research topics. To avoid such scenario, the globalization policies should not only support publishing in journals abroad, but also support local researchers in active participation in management of international journals and help correcting the "western bias" of journals.

The globalization efforts should reflect that the needs of national innovation systems in developing countries can be different than in western countries. Developing countries often aim to trigger imitation and diffusion of already existing technologies. Research policies often focus on increasing absorptive capacities of technological sector through i) creation of links with international sources and spotting feasible avenues for technological development and ii) triggering creation of interactive links between science and technology sector (Bernardes and Motta e Albuquerque, 2003). Blind adoption of globalization policies from advanced countries where policies pursue different goals and face different constraints can lead to even more serious issues of divergence of researchers' attention from the problems in their countries to attempts to comply with "international standards". If these attempts were unsuccessful, it might even lead to useless gaming instead of creating opportunities for development. Any such indications must be put under careful scrutiny.

References

Adams, J. (2012). Collaborations: The rise of research networks. In Nature (Vol. 490, Issue 7420, pp. 335–336). Nature Publishing Group. https://doi.org/10.1038/490335a

Adams, J. (2013). Collaborations: The fourth age of research. In Nature (Vol. 497, Issue 7451, pp. 557–560). Nature Publishing Group. https://doi.org/10.1038/497557a

Aman, V. (2016). Measuring internationality without bias against the periphery. 21st International Conference on Science and Technology Indicators, 1042–1050. https://inis.iaea.org/search/search.aspx?orig_q=RN:48050890

Auranen, O., & Nieminen, M. (2010). University research funding and publication performance-An international comparison. Research Policy, 39(6), 822–834. https://doi.org/10.1016/j.respol.2010.03.003

Bekavac, A., Petrak, J., & Buneta, Z. (1994). Citation behavior and place of publication in the authors from the scientific periphery: A matter of quality? Information Processing and Management, 30(1), 33–42. https://doi.org/10.1016/0306-4573(94)90022-1

Bernardes, A. T., & e Albuquerque, E. D. M. (2003). Cross-over, thresholds, and interactions between science and technology: lessons for less-developed countries. Research policy, 32(5), 865-885.

Bornmann, L., Wagner, C., & Leydesdorff, L. (2015). BRICS countries and scientific excellence: A bibliometric analysis of most frequently cited papers. Journal of the Association for Information Science and Technology, 66(7), 1507–1513. https://doi.org/10.1002/asi.23333

Brasil, A. (2021). Beyond the Web of Science: an overview of Brazilian papers indexed by regionally relevant databases. 18th International Conference on Scientometrics & Informetrics, 193–204. https://issi2021.org/proceedings/

Buela-Casal, G., Perakakis, P., Taylor, M., & Checa, P. (2006). Measuring internationality: Reflections and perspectives on academic journals. Scientometrics, 67(1), 45–65. https://doi.org/10.1007/s11192-006-0050-z

Chavarro, D., Tang, P., & Rafols, I. (2014). Interdisciplinarity and research on local issues: evidence from a developing country. Research Evaluation, 23(3), 195–209. https://doi.org/10.1093/reseval/rvu012

Chavarro, D., Tang, P., & Ràfols, I. (2017). Why researchers publish in non-mainstream journals: Training, knowledge bridging, and gap filling. Research Policy, 46(9), 1666–1680. https://doi.org/10.1016/j.respol.2017.08.002

Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'?. Research evaluation, 27(2), 106-118.

Ciarli, T., & Ràfols, I. (2019). The relation between research priorities and societal demands: The case of rice. Research Policy, 48(4), 949–967. https://doi.org/10.1016/j.respol.2018.10.027

Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. Applied Geography, 96, 1–10. https://doi.org/10.1016/j.apgeog.2018.04.017

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use-a literature review. In Research Evaluation (Vol. 25, Issue 2, pp. 161–169). Oxford University Press. https://doi.org/10.1093/reseval/rvv038

Franzoni, C., Scellato, G., & Stephan, P. (2011). Changing incentives to publish. Science, 333(6043), 702–703. https://doi.org/10.1126/science.1197286

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics, 3(3), 222–232. https://doi.org/10.1016/j.joi.2009.03.005

Gazni, A. (2015). Globalization of national journals: investigating the growth of international authorship. Learned Publishing, 28(3), 195–204. https://doi.org/10.1087/20150305

Gazni, A., Ghaseminik, Z., Gazni, A., & Ghaseminik, Z. (2016). Internationalization of scientific publishing over time: Analysing publishers and fields differences. Learned Publishing, 29(2), 103–111. https://doi.org/10.1002/LEAP.1018

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the American Society for Information Science and Technology, 63(2), 323–335. https://doi.org/10.1002/asi.21688

Good, B., Vermeulen, N., Tiefenthaler, B., & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. Research Evaluation, 24(2), 91–105. https://doi.org/10.1093/reseval/rvu035

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. Geoforum, 105, 1–12. https://doi.org/10.1016/j.geoforum.2019.06.017

He, T., & Liu, W. (2009). The internationalization of Chinese scientific journals: A quantitative comparison of three chemical journals from China, England and Japan. Scientometrics 2009 80:3, 80(3), 583–593. https://doi.org/10.1007/S11192-008-2067-Y

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

Hicks, D., & Wang, J. (2011). Coverage and overlap of the new social sciences and humanities journal lists. Journal of the American Society for Information Science and Technology, 62(2), 284–294. https://doi.org/10.1002/asi.21458

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. Research Policy, 39(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

IMF. (2003). The World Economic Outlook - Statistical Appendix. In The World Economic Outlook - April 2003 (pp. 163–169). https://www.imf.org/en/Publications/WEO/Issues/2016/12/31/Growth-and-Institutions

Kirchik, O., Gingras, Y., & Larivière, V. (2012). Changes in publication languages and citation practices and their effect on the scientific impact of Russian science (1993-2010). In Journal of the American Society for Information Science and Technology (Vol. 63, Issue 7, pp. 1411–1419). John Wiley & Sons, Ltd. https://doi.org/10.1002/asi.22642

Kuzhabekova, A., & Lee, J. (2018). International Faculty Contribution to Local Research Capacity Building: A View from Publication Data. Higher Education Policy, 31(3), 423–446. https://doi.org/10.1057/s41307-017-0067-3

Kuzhabekova, A., & Lee, J. T. (2020). Internationalization and Local Research Capacity Strengthening: Factors Affecting Knowledge Sharing Between International and Local Faculty in Kazakhstan. European Education, 52(4), 297–311. https://doi.org/10.1080/10564934.2020.1723422

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. PLoS ONE, 10(6), e0127502. https://doi.org/10.1371/journal.pone.0127502

Leta, J. (2012). Brazilian Growth in the Mainstream Science: The Role of Human Resources and National Journals. Journal Scientometric Research, 1(1), 44–52. https://doi.org/10.5530/jscires.2012.1.9

López Piñeiro, C., & Hicks, D. (2015). Reception of Spanish sociology by domestic and foreign audiences differs and has consequences for evaluation. Research Evaluation, 24(1), 78–89. https://doi.org/10.1093/reseval/rvu030

Ma, Z. (2019). The relevance of national journals from a Chinese perspective. In Springer Handbooks (pp. 505–562). Springer. https://doi.org/10.1007/978-3-030-02511-3_20

Macháček, V., Srholec, M., Ferreira, M. R., Robinson-Garcia, N., & Costas, R. (2021). Researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization. Science and Public Policy, 717–728. https://doi.org/10.1093/SCIPOL/SCAB064

Mattsson, P., Laget, P., Nilsson, A., & Sundberg, C. J. (2008). Intra-EU vs. extra-EU scientific co-publication patterns in EU. Scientometrics, 75(3), 555–574. https://doi.org/10.1007/s11192-007-1793-x

Moed, H. F. (2002). Measuring China's research performance using the science citation index. Scientometrics, 53(3), 281–296. https://doi.org/10.1023/A:1014812810602

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Pajić, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? Scientometrics, 102(3), 2131–2150. https://doi.org/10.1007/s11192-014-1510-5

Ponds, R. (2009). The limits to internationalization of scientific research collaboration. The Journal of Technology Transfer, 34, 76–94. https://doi.org/10.1007/s10961-008-9083-1

Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). Aslib Journal of Information Management, 69(5), 486–502. https://doi.org/10.1108/AJIM-01-2017-0014

Rafols, I., Molas-Galart, J., Chavarro, D., & Robinson-Garcia, N. (2016). On the Dominance of Quantitative Evaluation in Peripherall Countries: Auditing Research with Technologies of Distance. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2818335

Ren, S., & Rousseau, R. (2002). International visibility of Chinese scientific journals. Scientometrics 2002 53:3, 53(3), 389–405. https://doi.org/10.1023/A:1014877130166

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. Journal of Informetrics, 13(1), 50–63. https://doi.org/10.1016/j.joi.2018.11.002

Royal Society. (2011). Knowledge, networks and nations: Global scientific collaboration in the 21st century. https://royalsociety.org/-/media/Royal_Society_Content/policy/publications/2011/4294976134.pdf

Science Europe. (2013). Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility. https://www.elsevier.com/__data/assets/pdf_file/0019/53074/Comparative-Benchmarking-of-European-and-US-Research-Collaboration-and-Researcher-Mobility_sept2013.pdf

Scopus. (2018). Scopus Source List (May 2018) version. https://www.scopus.com/sources

Scopus. (2019). What is the complete list of Scopus Subject Areas and All Science Journal Classification Codes (ASJC)? https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/related/1/

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports 2019 9:1, 9(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z

Wagner, C., Brahmakulam, I., Jackson, B., & Wong, A. (2001). Science and Technology Collaboration: Building Capacity in Developing Countries. https://www.researchgate.net/profile/Caroline-Wagner-2/publication/238311644_Science_Technology_Collaboration_Building_Capacity_in_Developing_Countries_The_RAND_Corporation/links/582b1fa208ae102f07207868/Science-Technology-Collaboration-Building-Capacity-in-Developing-Countries-The-RAND-Corporation.pdf

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. In Nature (Vol. 550, Issue 7674, p. 33). Nature Publishing Group. https://doi.org/10.1038/550032a

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The Continuing Growth of Global Cooperation Networks in Research: A Conundrum for National Governments. PLOS ONE, 10(7), e0131816. https://doi.org/10.1371/journal.pone.0131816

Wagner, C. S., Whetsell, T. A., & Leydesdorff, L. (2017). Growth of international collaboration in science: revisiting six specialties. Scientometrics, 110(3), 1633–1652. https://doi.org/10.1007/s11192-016-2230-9

Wagner, C. S., & Wong, S. K. (2012). Unseen science? Representation of BRICs in global science. Scientometrics, 90(3), 1001–1013. https://doi.org/10.1007/S11192-011-0481-Z/TABLES/3

Waltman, L., Tijssen, R. J. W., & Eck, N. J. van. (2011). Globalisation of science in kilometres. Journal of Informetrics, 5(4), 574–582. https://doi.org/10.1016/j.joi.2011.05.003

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. Journal of Informetrics, 10(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Wang, S., Wang, H., & Weldon, P. R. (2007). Bibliometric analysis of English-language academic journals of China and their internationalization. Scientometrics 2007 73:3, 73(3), 331–343. https://doi.org/10.1007/S11192-007-1775-Z

Zhang, L., Shang, Y., Huang, Y., & Sivertsen, G. (2021). Toward internationalization: A bibliometric analysis of the social sciences in Mainland China from 1979 to 2018. Quantitative Science Studies, 2(1), 376–408. https://doi.org/10.1162/qss_a_00102

Zhou, P., & Glänzel, W. (2010). In-depth analysis on China's international cooperation in science. Scientometrics, 82(3), 597–612. https://doi.org/10.1007/s11192-010-0174-z

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. Research Policy, 35(1), 83–104. https://doi.org/10.1016/j.respol.2005.08.006

Zitt, M., & Bassecoulard, E. (1998). Internationalization of scientific journals: A measurement based on publication and citation scope. Scientometrics, 41(1–2), 255–271. https://doi.org/10.1007/BF02457982

Zitt, M., & Bassecoulard, E. (1999). Internationalization of communication a view on the evolution of scientific journals. Scientometrics, 46(3), 669–685. https://doi.org/10.1007/BF02459619

Zitt, M., & Bassecoulard, E. (2004). Internationalisation in Science in the Prism of Bibliometric Indicators. In Handbook of Quantitative Science and Technology Research (pp. 407–436). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_19

Zitt, M., Perrot, F., & Barré, R. (1998). The transition from "national" to "transnational" model and related measures of countries' performance. Journal of the American Society for Information Science, 49(1), 30–42. https://doi.org/10.1002/(SICI)1097-4571(1998)49:1<30::AID-ASI5>3.0.CO;2-3

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2003). Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. Scientometrics 2003 56:2, 56(2), 259–282. https://doi.org/10.1023/A:1021923329277

Appendix

Supplementary file 3.1: Mean globalization scores for all countries, disciplines, years and indicators

https://github.com/vitekzkytek/GlobalizationPaper/blob/master/public_data/globalization_scores.csv

Supplementary file 3.2: Distribution of research output into journals by globalization (Euclidian distance) in broad disciplines

https://github.com/vitekzkytek/GlobalizationPaper/blob/master/public_data/globalization_quartiles.csv

Supplementary file Appendix A4: Individual journal globalizations (Euclidian distance, broad set of disciplines, 2017)

Note that journals with multiple disciplines are computed for each discipline. The results will differ for journals with multiple disciplines for benchmark-based indicators.

https://github.com/vitekzkytek/GlobalizationPaper/blob/master/public_data/globalization_journals.csv

Appendix Table 3.3: Classification of countries

Category	N	List
Advanced countries	32	Australia, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany, Greece, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Singapore, South Korea, Spain, Sweden, Switzerland, Taiwan, United Kingdom, United States, Liechtenstein, Monaco, Malta
Developing – Africa	49	Algeria, Benin, Botswana, Burkina Faso, Cameroon, Congo, Cote d'Ivoire, Egypt, Ethiopia, Gabon, Ghana, Kenya, Madagascar, Malawi, Mali, Morocco, Mozambique, Namibia, Nigeria, Saudi Arabia, Senegal, South Africa, Sudan, Tanzania, Tunisia, Uganda, Zambia, Zimbabwe, Libya, Gambia, Mauritius, Niger, Togo, Eritrea, Guinea, Rwanda, Swaziland, Lesotho, Angola, Democratic Republic Congo, Sierra Leone, Central African Republic, Seychelles, Mauritania, Guinea-Bissau, Burundi, Liberia, Cape Verde, Chad
Developing – Asia and Pacific	35	Bangladesh, Cambodia, China, India, Indonesia, Iran, Jordan, Kuwait, Lebanon, Malaysia, Nepal, Oman, Pakistan, Philippines, Sri Lanka, Syria, Thailand, Turkey, United Arab Emirates, Vietnam, Bahrain, Fiji, Iraq, North Korea, Palestine, Qatar, Brunei, Laos, Myanmar, Papua New Guinea, Yemen, Afghanistan, Bhutan, Vanuatu, Solomon Islands

Developing - America	30	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Guatemala, Jamaica, Mexico, Panama, Peru, Trinidad and Tobago, Uruguay, Venezuela, Barbados, El Salvador, Honduras, Nicaragua, Paraguay, Dominican Republic, Grenada, Haiti, Bahamas, Saint Kitts and Nevis, Dominica, Guyana, Suriname, Belize
Transition - EU	11	Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia
Transition - non-EU	17	Armenia, Belarus, Bosnia and Herzegovina, Georgia, Kazakhstan, Macedonia, Mongolia, Russia, Ukraine, Uzbekistan, Azerbaijan, Kyrgyzstan, Moldova, Serbia, Albania, Tajikistan, Montenegro

Appendix: Response to reviewers' comments

Predatory journals

Vincent Larivère

- *Reviewer notes that analysis of micro-level data would clarify the individual characteristics of researchers, especially their seniority.*

  This is certainly true and the reviewer is correct that we did not have such data available. Paragraph was added to conclusions suggesting this future stream of research (see p. 50).

- *Reviewer also considers a possibility to include a variable related to research evaluation culture into a regression.*

  We have spent quite time finding an operationalization that would allow exactly that, even beyond the research evaluation landscape, by operationalizing quality of governance in general. However, the broad country scope of the analysis did not allow to do it. Paragraph explaining this already was in the conclusions: "we would like to take characteristics of the research evaluation framework directly into account, including whether evaluation primarily concerns quantity or quality, whether formulae based on quantitative metrics is used, how advanced the underlying bibliometric approach is, whether insights from peer review assessment are factored in, and, consequently, what principles are applied when allocating research funding. Unfortunately, indicators of this kind are not available for more than a handful of advanced countries, which are not the most relevant here" (p. 49 – 50).

- *The reviewer contests a statement „The more the research evaluation system relies on outdated routines such as counting articles indexed in Scopus, Web of Science or Medline, the higher incentive for researchers to publish in fraudulent journals just to clinch points for outputs regardless of merit" (p. 23). He emphasizes a strong filtering function that bibliometric databases have and defends use of bibliometric measures as they seem not to be providing a fruitful ground for predatory journals. Countries with strong reliance on bibliometrics-based evaluation arguably do not have much predatory publishing. Additionally, he names even worse evaluation practice such as counting papers in CVs (i.e. evaluation without any kind of filtering).*

  Regarding the use of bibliometric measures – the above "outdated routines such as counting articles" does not challenge using bibliometrics-based evaluation practice in general. It is really focused on evaluation based on simplistic cash-per-paper policies.

  Perhaps the most important point of this paper is an evidence that relying on bibliometric indexation filtering is not sufficient as some problematic journals can find their ways into it. Although 324 journals does not seem to constitute a big issue (relative to 30 000 journals indexed in Scopus), but in 2015 it was approximately 3% of all publications indexed in Scopus (see Macháček and Srholec, 2017). Especially, when looking into some of hardest hit countries these figures even reach 10 % and more. Arguably, the blind reliance on indexation in bibliometric databases, hand in hand with otherwise weak evaluation culture can boost the demand for predatory publishing. I have rephrased the relevant sentence to increase clarity (p. 23).

- *Reviewer identified a minor mistake in the used name of Frontiers publisher. The name we refer to - Frontiers Media Foundation – is incorrect as the Frontiers Foundation is a different entity.*

We derived this name from the Ulrichsweb database during data gathering processes, but will use more neutral Frontiers publisher instead.

- *Another minor mistake is in a sentence "Most Frontiers journals are also indexed in the Web of Science and the Directory of Open Access Journals. (p. 27)"*

  Reviewer is correct that many Frontiers journals are not indexed in WoS. The sentence to was rephrased to "*Many journals ...*"

Daniel Münich

- *Text does not describe delta coefficient properly in Equation 1 (p. 43).*

  Fixed in the manuscript, sorry for a typo

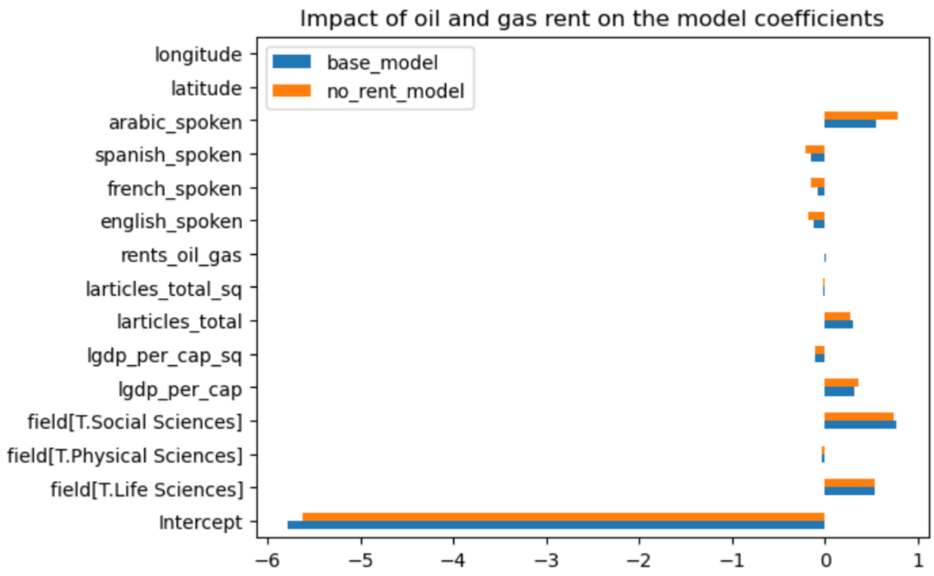
- *Reviewer suggests using number of authors instead of number of documents as a proxy for research sector size. He is afraid that number of documents is inflated by the number of predatory publications and therefore might be distorting the results.*

  The argument makes sense as the predatory articles might have a hard time to get published in other Scopus-indexed venues than predatory journals. Unfortunately, our data do not contain this variable (due to the downloading procedure).

  However, the distortion resulting from the use of number of documents is likely to be rather small. The differences between countries in terms of research sector size are large – for example China had more than 13 times more relevant documents than Sweden in respective period. The inflation induced by the using number of documents is however limited by the estimated propensities to publish in predatory journals – in a few countries over 15 % and in most countries below 5 %.

- *Reviewer asks for better discussion of oil and natural gas coefficient*

  The coefficient is stable around 0.02 in most models specifications. At median this means that increasing the rent by 1pp of GDP is associated with 2 % increase in the tendency to publish in predatory journals – i.e. the coefficient is large. However, interpreting the particular coefficient can be troublesome due to highly skewed distribution of the underlying data – only a handful of countries have high percentage of GDP. It is sufficient to generally conclude that higher oil and gas rent can contribute to the tendency to publish in predatory journals. The variable is not essential for the model. The models' coefficients remain stable even when the variable is not considered at all.



- *Reviewer asks for better discussion of proportion of the variance explained by explanatory variables as well as only of across fields and he is interested in the effect size for fields*

1. Proportion of variance by explanatory variables is not easy to deliver in the GLM framework, as the classical R2 has no straightforward calculation. For simplicity we decided to re-estimate the model using standard OLS, solely for the purpose of R2 estimation. The R2 is approximately 19 % for Total column in Fig 1.5 as well as for column Total excluding Frontiers.
2. To deliver proportion of variance explained by field dummies we followed similar procedure and estimated standard OLS model with only 1 variable – field. This estimation produce approximately 5 % R2 for dependent variables "Total" and "Total excluding Frontiers". Fields do play a role in explaining the model, however it is not dominant driver of explained variance.
3. When we use Health sciences as a reference, the estimated effects of field dummies are: a) 0.54 in Life sciences, -0.035 in Physical sciences and 0.77 for Social sciences. However, in all three the standard errors are wide so it is not advisable to interpret it at all.

- *Two explanatory variables are inserted in quadratic form. Did you explore nonlinear impact of the latitude and longitude too?*

The nonlinear properties of geography was not studied as (similarly as in the case of oil and gas rent) we were interested in factors at least vaguely related to the research environment in the country. We decided to include latitude and longitude into the analysis to account for general North-South and East-West divide, however I do not see any value added from allowing non-linearity that would not be evident from the Figure 1.1 (the map). The u-shaped properties might point towards middle east as a centre of gravity in both west-east dimension and in north-south dimension. However, this effect is already included in oil and gas rent as well as in Arabic language.

Thed van Leeuwen

- *The reviewer notes that perhaps the most important contribution of this paper is the publications aftermath including an unjustified retraction from Scientometrics and the re-publication in Quantitative Science Studies.*

  I agree that the attention dedicated to the papers' retraction triggered more attention than the content itself. However, I would hesitate to mark it as a contribution of my dissertation.

Researchers' institutional mobility

Vincent Larivère

- *The reviewer suggests addressing the country language as a factor explaining reliance on domestic resources in the university hiring practices. Universities in countries with rather small number of speakers might be forced to rely from domestic resources just based on this simple argument of "availability of resources".*

  This was a great suggestion to resolve, as the reviewer is right that our analysis so far neglected the role of socio-economic factors that are beyond the control of universities and of the research system in general. We agree that the role of language as a factor influencing institutional mobility could shed more light into the analysis.

  We extended the analysis to account for the number of speakers of languages spoken in the given country. This analysis is appended to the end of results section (p. 78 – 80). Unfortunately, the results of this preliminary analysis do not support the hypothesis that number of speakers is significant factor for the share of insiders.

Daniel Münich

- *Reviewer identified a typo. It should read "Our analysis explores…".*

  Corrected (p. 67) . Thank you for the notice.

- *Table 2.1: What is the average "Number of researchers"? Is it a mean per university?*

  Yes, it is, was clarified in the table caption.

- *Reviewer asked for more detailed zoom on regions of Europe and North America.*

  We have extended the Figure 2.1 on the full page, so that the more detail is visible. Also, zoomed maps to Europe and North America were added to the appendix (p. 92).

- *Unclear description of table 2.3 (variance decomposition)*

  The manuscript was slightly rephrased to increase clarity.

Thed van Leeuwen

- *The reviewer compares our operationalization of researchers to Gaye Tuchman's book, 'Wannabe U. Inside the Corporate University'.*

  It is great that reading the paper triggers thoughts beyond standard problems of research evaluation – such as the corporatization and how pressure on generation of "measurable outcomes" may hinder development. It would be great if someone used our methodology to get more informed insight on effects of corporatization.

- *The reviewers asks to what extent the globalization and mobility are determined institutionally and less so they are determined by individual decisions of scholars.*

  I agree this is fundamental question underlying the dissertation, however the dissertation provides only limited understanding to this besides what reviewer has already noted. I believe this is a viable topic for future research and the data and methods developed in this dissertation can be used to start answering these question. But honestly, this could be a content of an entirely new dissertation. One new paragraph in the conclusions actually describe the multi-level nature of factors determining mobility decisions (p. 85).

Globalization of scientific communication

Vincent Larivère

- *The reviewer asks for better discussion of the conclusions related to the relationship of globalization of scientific communication and "westernization of research".*

  The last paragraph mentioning the paper about rice was rephrased and extended to account for more general implications that westernization of research can have on developing countries. Most important implication is that globalization is only a tool to achieve some more general research policy goals (p. 119 - 120).

- *Discussion of the limitations of the use of the Scopus database to measure those westernization aspects*

  I believe that besides the alignment of globalization with research policy goals described above, the major limitation of the Scopus database is representativeness – an issue that is well described both in the Data section (p. 106 - 107) as well as in the conclusions (p. 118, "It cannot be stressed enough …")

Daniel Münich

- *Reviewer asks for changing the map projection to increase readability of Europe relative to relatively homogenous Africa.*

  The projection was changed to EckertIII and perhaps more importantly the figure itself was extended to the full page to increase readability (p. 109).

- *The reviewer notes that in Figure 3.4 there is Czech Republic missing Czech Republic*

  Thank you for pointing this out. There was a small typo in the code "Czechia" instead of "Czech Republic". Sorry for that, the figure was fixed in the final manuscript (p. 113).

Thed van Leeuwen

- *The reviewer acknowledges the universalistic vs. contextual distinction and according to him it can help in understanding the differences.*

  This distinction tries to explain the differences between social and health sciences and the physical and life sciences in some countries (for example in Central and Eastern Europe).

However, these differences are not everywhere. Moreover, this theory was not tested at all and whether this actually is what makes a difference. Test this assumption beyond the scope of this analysis, but I believe it could be done in some kind of NLP framework.

After that even more interesting questions will occur. What kind of research do we actually want to do in SSH – is contextual better than universalistic, or vice versa? Do we want them both? Do we want them published in different venues, or it would be better if both approaches would occur in the same globalized journals. I truly think that this is the beginning of the story and we should go beyond this rather simplistic distinction.

References

Macháček, V. and Srholec, M. (2017). Predatory journal in Scopus. IDEA study 2 / 2017, https://idea-en.cerge-ei.cz/files/IDEA_Study_2_2017_Predatory_journals_in_Scopus/mobile/index.html#p=1.