D3.2 Cumulative Advantage in Open Science and RRI: A Large-Scale Quantitative Study

Observing and Negating Matthew Effects in Responsible Research and Innovation Transition

D3.2 Cumulative Advantage in Open Science and RRI: A Large-Scale Quantitative Study
Version 1.0
Public

This document reports on the research activities conducted to identify, measure and assess cumulative advantage in Open Science and RRI with a clear focus on the creation of research outputs within academia. We process several large scholarly datasets and conduct four original quantitative research studies exploring how OS and RRI practices, compared across geographical and institutional boundaries, are associated with institutional rankings, GDP, gender and other demographic data. While the first two research studies analyse all available data across the globe, providing the broadest possible macro view, the other two studies offer an in-depth investigation of the situation within three target domains defined by the UN Sustainable Development Goals (SDGs) - SDG 2 - Zero Hunger, SDG 3 - Good Health and Well-Being and SDG 13 - Climate.

ON-MERRIT - Grant Agreement 824612

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612.

Document Description

D3.2 Cumulative Advantage in Open Science and RRI: A large-scale quantitative study.

DX.X Name			
WP3 - Research cultures, support and incentives			
Due date	30/09/2021	Actual delivery date:	30/09/2021
Nature of document	[Report]	Version	1.0
Dissemination level	Public		
Lead Partner for deliverable	The Open University		
Authors	David Pride, Thomas Klebel, Matteo Cancellieri, Tony Ross-Hellauer, Petr Knoth		
Reviewers	Angela Fessl, Ilaria Fava, Birgit Schmidt		

Revision History

Issue	Item	Comments	Author/Reviewer
V 0.1	Draft version		DP, TK, PK
V 0.2	Revised	Revision based on feedback from the internal review and ON-MERRIT Advisory Board.	DP, TK, TRH, MC, PK
V 1.0	Revised	Final revisions	DP, TK, TRH, PK

Table of Contents

Tables

Figures

Abbreviations

CWTS - Centre for Science and Technology Studies
DOAJ - Directory of Open Access Journals
EC – European Commission
GDP - Gross Domestic Product
HIC - High income countries
LIC - Low income countries
LMIC - Lower middle income countries
MAG - Microsoft Academic Graph
MIC - Middle income countries
OA - Open Access
OS - Open Science
RRI - Responsible Research and Innovation
SDG - Sustainable Development Goal
THE - Times Higher Education
THE WUR - Times Higher Education World University Rankings
UMIC - Upper middle income countries
WoS - Web of Science
WP – Work Package

Executive summary

This document reports on the research activities conducted in "ON-MERRIT's" Work Package 3 to identify, measure and assess effects of cumulative advantage in Open Science and RRI with a clear focus on the creation of research outputs within academia. We conduct four original quantitative research studies addressing a range of pertinent research questions, including: *Who produces and who consumes open access research literature? How is institutional performance related to the application of RRI policies and OA publishing? Does the uptake of OA publishing change existing hierarchies within academic publishing, and if so, in which ways across a subset of ON-MERRIT's target UN Sustainable Development Goals (SDGs) - SDG 2 - Zero Hunger, SDG 3 - Good Health and Well-Being and SDG 13 - Climate Action?*

The first research study investigates levels of production and consumption of Open Access (OA) research literature globally, measured as the proportion of citations to OA literature, and tests for correlations with gross domestic product (GDP) per capita at the country and continental level. We find moderate correlations between OA production and OA consumption. We find a stronger correlation for higher ranked institutions when using ranking data from the Times Higher Education (THE) World University Rankings (WUR). We surprisingly find no correlation between OA production and consumption and GDP (per capita).

The second research study investigates how performance or prestige, primarily at the institutional level, is related to the application of RRI and OA publishing. We observe that the most highly ranked institutions are both greater producers and greater consumers of OA. We find a strong overall correlation between *public engagement* with science, and RRI policies at the national level. A further interesting final finding is the lack of correlation between how a country performs in terms of gender equality policies and the balance in ratios of male / female researchers. This is particularly noticeable in the new EU13 countries.

Following ON-MERRIT's focus on three key UN SDGs (SDG Zero Hunger (SDG 2), SDG Health/Well-Being (SDG 3) and SDG Climate Action (SDG 13)), the third research study surveys the extent of growth and impact of this literature across the analysed SDGs, as well as its structure in terms of who contributes. We find the literature on SDGs to be increasing by 5-7% per year on average, with an increase in the share of female authorships (27-37% in 2019, depending on the SDG), and persistent institutional stratification.

In the fourth study, we continue our investigation of literature related to UN SDGs, focusing on aspects of OA publishing. We analyse how the uptake of OA publishing differs according to dimensions such as gender, academic age, institutional prestige, and country income. We find that well-resourced actors publish more frequently OA in the SDG areas, as well as publishing in journals with on average higher APCs, which might worsen already existing structural hierarchies within academia.

The four studies presented in this deliverable combine to highlight that it is the higher ranked, more prosperous and more prestigious institutions that appear best able to adopt, adapt to, and benefit from, the evolving landscape of Open Access publishing. These trends hold true over time, on the global level, and when broken down to individual continents and subject areas (SDGs). Persistent structural inequalities in contemporary academic publishing are not necessarily remedied by the Open Science movement, with specific trends such as APC-driven OA publishing potentially exacerbating dynamics of cumulative advantage. If research on key global issues is only driven by well-resourced actors, it risks being oblivious to challenges faced by societies and communities less embedded into the global production of knowledge.

1. Introduction

This document reports on the research activities conducted to identify, measure and assess cumulative advantage in Open Science and RRI with a clear focus on the creation of research outputs within academia. We rely on several large scholarly datasets to conduct four original quantitative research studies addressing a range of pertinent research questions, including: *Who produces and who consumes open access research literature? How is institutional performance related to the application of RRI policies and OA publishing? Does the uptake of OA publishing change existing hierarchies within academic publishing?, and if so, in which ways across a subset of ON-MERRIT's target UN Sustainable Development Goals (SDGs) - SDG 2 - Zero Hunger, SDG 3 - Good Health and Well-Being and SDG 13 - Climate Action?*

Our work is motivated by the "Matthew effect" phenomenon. The term 'Matthew effect' was originally conceived by (Merton 1968). Merton's definition of the Matthew Effect was fairly narrow, referring specifically to the mis-allocation of credit for scientific work. However, the underlying theme of the work was that this mis-allocation was just one of many examples that can be found in academia of 'the rich getting richer'. In a virtuous (though some may say vicious) cycle, a highly cited author becomes even more highly cited, not due to any intrinsic quality in their work, but as a by-product of the author's existing status. Further, highly prestigious institutions are able to attract more funding and more students, which in turn allows for more investment in resources which then continues this cycle. De Solla Price (1965) coined the term *cumulative advantage* to refer to these effects. It was demonstrated that these effects are visible for highly-regarded scientists, institutions, and even journals.

The four research studies presented here are all connected and motivated by the notion of cumulative advantage and investigate different segments of it. While studies one and two focus on the macro aspects delivering large-scale and broad analyses; studies three and four look more into depth providing valuable insights into the dynamics within three analysed domains defined by SDGs.

We investigate aspects of cumulative advantage along a range of dimensions including geographical location, gender and institutional standing. The work builds on ON-MERRIT D3.1[1] which created the initial datasets which were later refined to conduct these studies.

For the studies described in this report, we use a range of publicly available data sources. Publication metadata for research papers and institutions is retrieved from Microsoft Academic Graph (MAG). The Open Access (OA) status for these research papers is retrieved using Unpaywall[2]. We use two separate, internationally recognised, ranking indicators: the Times Higher Education (THE) World University Rankings (WUR) and the Leiden Ranking, developed and maintained by the Centre for Science and Technology Studies (CWTS). We also use rankings derived from MAG data as a comparison. We use economic data from the World Bank Indicators and data on Responsible Research and Innovation (RRI) from the EC-funded MoRRI project (2014-2018).[3]

The first research study investigates levels of production and consumption of Open Access research literature globally. It poses the question '*Who produces and who consumes open access research literature?*'.

[1] https://doi.org/10.5281/zenodo.3874586
[2] https://unpaywall.org/
[3] https://data.europa.eu/data/datasets/morri_data?locale=en

Answering this question helps us to identify and gain a better insight into the entities (continents, countries, institutions) that are the driving forces behind the rise of open access as well as to understand who are the main users benefiting from it. The study examines levels of OA production and OA consumption, measured as the proportion of citations to OA literature, globally and tests for correlations with gross domestic product (GDP) per capita at the country and continental level. This is to investigate whether the overall economic status of a particular country is associated with the magnitude of production and or use of open access. We find medium-strong correlations between OA production and OA consumption. We find a stronger correlation for the higher ranked institutions when using ranking data from the Times Higher Education (THE) World University Rankings (WUR). We surprisingly find no correlation between OA production and consumption and GDP (per capita).

Following this, the second research study investigates '*How is performance or prestige, primarily at the institutional level, related to the application of RRI and OA publishing*?'. Combining research paper metadata with RRI data and measures of institutional prestige / rank (such as the Times Higher rankings or Leiden Ranking) allows for a detailed investigation into the interplay between these different factors. We observe the most highly ranked institutions are both greater producers and greater consumers of OA than lower-ranked institutions, although the size of this difference can vary depending on the ranking used. We find a strong overall correlation between *public engagement* with science, and RRI policies at the national level. We see demonstrably higher levels of public engagement with science in countries where these policies are more embedded. Further, we show a medium to strong correlation between institutional prestige and OA production which aligns with our earlier findings. A further interesting finding is the lack of correlation between how a country performs in terms of gender equality policies and the balance in numbers of male / female researchers. This is particularly noticeable in the new EU13 countries. This study also helps to clarify who is benefitting from the application of RRI policies and mandates and can guide suggestions for how this can be improved in the future.

Following ON-MERRIT's focus on three key UN SDGs (SDG Zero Hunger (SDG 2), SDG Health/Well-Being (SDG 3) and SDG Climate Action (SDG 13)), the third research study surveys the extent of growth and impact of this literature across the analysed SDGs, as well as its structure in terms of who contributes. This investigation into factors such as gender, academic age and institutional prestige lays the groundwork for the next section of the study. We find the literature on SDGs to be increasing by 5-7% per year on average, with an increase in the share of female authorships (27-37% in 2019, depending on the SDG), and persistent institutional stratification.

In Study four, we continue our investigation of literature related to UN SDGs, focusing on aspects of OA publishing. Combining rich data from MAG, Unpaywall, the World Bank, the Leiden Ranking and the Directory of Open Access Journals (DOAJ), we show how the uptake of OA publishing differs according to dimensions such as gender, academic age, institutional prestige, and country income. Using data from DOAJ, we show how the OA publishing model advantages more resourceful actors, which might worsen already existing structural hierarchies within academia.

In the final section, we collect, synthesize and discuss all findings from the four substantial studies and provide recommendations. The four studies presented in this deliverable combine to highlight that it is the higher ranked, more prosperous and more prestigious institutions that appear best able to adopt, adapt to, and benefit from, the evolving landscape of Open Access publishing. These trends hold true over time, on

the global level, and when broken down to individual continents and subject areas (SDGs). Persistent structural inequalities in contemporary academic publishing are not necessarily remedied by the Open Science movement, with specific trends such as APC-driven OA publishing potentially exacerbating dynamics of cumulative advantage. If research on key global issues is only driven by well-resourced actors, it risks being oblivious to challenges faced by societies and communities less embedded into the global production of knowledge.

2. Data Sources

In the following section we introduce the data sources used for the experiments in the studies contained within this deliverable. These data sources fall into several distinct categories, notably research paper data from Microsoft Academic Graph and Unpaywall, university ranking data from Times Higher Education, The Leiden Rankings from CWTS and again, Microsoft Academic. We use economic data (Gross domestic product per capita) from the World Bank Indicators. And finally, we employ the Responsible Research and Innovation (RRI) data from the EC-Funded 'Monitoring the Evolution and Benefits of Responsible Research and Innovation' (MoRRI) project. A combination of these datasets are used for each study within this report, the rationale for the application of each particular dataset is given within the relevant study.

2.1. Microsoft Academic Graph

The Microsoft Academic Graph (MAG) was established in June 2015 and models '*real-life academic communication activities as a heterogeneous graph*' (Sinha et al. 2015a). MAG data can be accessed via the Microsoft Academic search engine[4] via the Academic Knowledge API or downloaded as a complete dataset. For these experiments a full download of the MAG dataset was made in November 2020.

Harzing and Alakangas (2017) demonstrate on a sample of 145 authors from 37 domains that the mean h-index of these authors is similar when using data from Microsoft Academic, Scopus or Web of Science; (MAG: 22; Scopus: 22, WoS: 20)

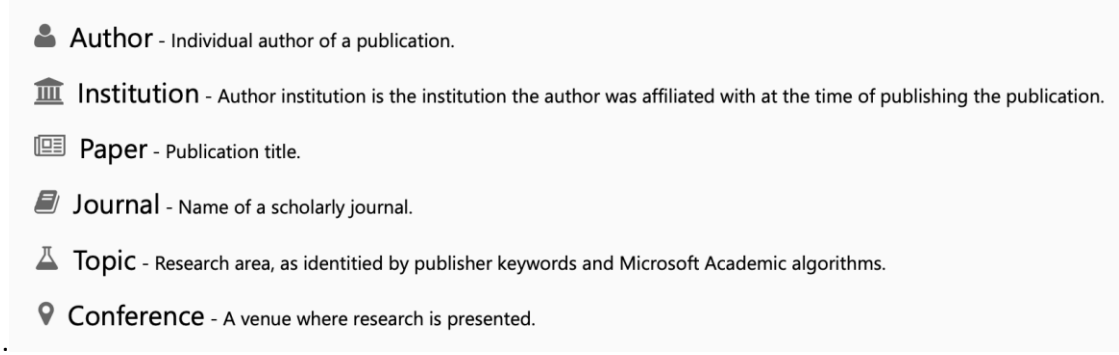

Figure 1: Microsoft Academic Graph entities

The data contained in MAG is obtained from web pages crawled by the Microsoft 'Bing' search engine. MAG is organised as a collection of entities using an XML style schema. Figure 1 shows the top level entity for each category. A full diagram of the MAG schema and all entities can be viewed online[5].

Studies have compared it to other bibliographic sources (Harzing and Alakangas 2017; Hug, Ochsner, and Brändle 2017) and found it to be an extremely comprehensive source of bibliometric data. However, they highlight some remaining concerns regarding the accuracy of the metadata. It should be said however that this issue is not only a problem of Microsoft, a recent study (Donner, Rimmert, and van Eck 2020) also highlighted issues regarding the affiliation disambiguation systems in WoS and Scopus.

[4] https://academic.microsoft.com

[5] https://docs.microsoft.com/en-us/academic-services/graph/media/erd/entity-relationship-diagram.png

It was announced by Microsoft in May 2021 that MAG would be retired at the end of the year.[6] As our analyses are conducted well in advance of MAG's retirement, this fortunately has no impact on the quality or timeliness of the data we use in our study. However, the loss of MAG will be very unfortunate for scholarly communication research. MAG is currently the second largest scholarly document search engine behind Google Scholar and its demise is regarded as a great loss by many in the bibliometric community. The unexpected decision to bring an end to MAG has galvanised many of the other players in the OA landscape to begin the process of scoping and building an equivalent service. [7]

2.2. The Leiden Ranking

The CWTS Leiden Ranking (Waltman et al. 2012) is produced by the Centre for Science and Technology Studies (CWTS) at Leiden University in the Netherlands. It currently offers a range of bibliometric performance indicators for approximately 1200 institutions worldwide. For all our analyses of institutional prestige, we rely on the 2020 Leiden Ranking (Van Eck 2021). This version of the Leiden Ranking includes 1,176 universities from 65 different countries. To be included, a given university has to have produced at least 800 Web of Science indexed publications in the period 2015–2018. For this, publications are counted fractionally (by dividing counts of authorships by the total number of authors on a given publication), and only research articles and review articles, and only those from so-called "core journals" are included. Indicators are standardized using algorithmically-generated micro-level fields (Waltman and van Eck 2012) and are calculated based on core publications. Core publications are publications that a) have been written in English, b) have one or more authors, c) have not been retracted, and d) are published by a core journal.

The motivation for using university rankings was in finding an indicator that ranks institutions according to their level of prestige within academia. From the available indicators, we chose to use $P_{top\ 10\%}$ for all our analyses. $P_{top\ 10\%}$ is defined as "*The number [...] of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 10% most frequently cited.*"[8] We chose the size-dependent $P_{top\ 10\%}$ over the size-independent $PP_{top\ 10\%}$ because previous research (Frenken, Heimeriks, and Hoekman 2017) has emphasised the role of university age and size when it comes to the level of resources available for supporting research activities (through research equipment, graduate programmes, libraries, institutional assistance in securing grant funding, etc.). Furthermore, we conducted many of the analyses with both $P_{top\ 10\%}$ and $PP_{top\ 10\%}$, and results were very similar overall.

Besides $P_{top\ 10\%,}$ we also used OA statistics for individual institutions which are present in the Leiden Ranking (Robinson-Garcia, Costas, and Leeuwen 2020). As we intended to go beyond institutional aggregates of OA publishing, we matched institutions from the Leiden Ranking to institutions from MAG. A first match on university names was successful for 945 out of 1176 institutions (80.4%). We manually tried to match the remaining institutions from the Leiden Ranking to MAG affiliations, using information from both datasets, as well as the MAG website, the Leiden Ranking Website, as well as Wikipedia. We were unable to match 14 universities, thus resulting in 98.8% matched institutions in total. An important limitation in this approach is that some definitions (e.g. what constitutes a university, which branches belong to one entity or are separate entities, etc.) between MAG and the Leiden Ranking (which is based on the Web of Science) might differ.

[6] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/
7 https://www.natureindex.com/news-blog/microsoft-academic-graph-discontinued-whats-next
8 https://www.leidenranking.com/information/indicators

2.3. Times Higher Education - World University Rankings

Times Higher Education (THE) World University Rankings (WUR) includes more than 1,500 institutions across 93 countries in 2021. THE state that these rankings, '*use 13 carefully calibrated performance indicators to provide the most comprehensive and balanced comparisons between institutions.*'[9] The performance indicators are grouped into five areas: Teaching (the learning environment); Research (volume, income and reputation); Citations (research influence); International outlook (staff, students and research); and industry income (knowledge transfer). The ranking for 2022 analysed more than 108 million citations across over 14.4 million research publications and included survey responses from almost 22,000 academics globally, and 2,100 institutions submitted data. The knowledge of exact data sources for the composition of the rankings, and the specific algorithms used to calculate them are the proprietary intellectual property of The Times organisation.

2.3.1. A comparison of ranking methodologies used

It should be stated that the usage of rankings, in whatever form they may take, when used to rank universities is controversial. Brankovic (2021) notes that any ranking system with a closed number of participants is a zero-sum game. One institution can only benefit at the expense of another. This may be viewed as illogical if one accepts the notion that one institution can clearly both improve, or worsen, entirely independently of the performance of another institution. Additionally, the use of closed or proprietary data, such as in the WUR rankings, is viewed as problematic. Gadd (2021) also covers a wide range of issues with current ranking systems such as gaming, 'industrialised self-citation' and selective submissions amongst other problems.

The Times Higher Education (THE) THE World University Rankings (WUR) aspires to "*provide the definitive list of the world's best universities*"[10]. There are however other world rankings, noticeably the Shanghai Academic Ranking of World Universities (ARWU)[11] and the QS rankings[12] that also have the same goal, and make similar claims. When one looks at the actual rankings however, they are clearly not identical and this raises the question of the efficacy of these and other methodologies used to rank universities. As noted by Waltman and van Eck (2012):

> *These composite indicators combine different dimensions of university performance in a rather arbitrary way. The fundamental problem of these indicators is the poorly defined concept of university performance on which they are based.*

We argue that a composite indicator that conflates research performance and teaching performance does justice to no one. It is entirely possible that an institution may be exemplary in teaching undergraduates yet have far less established or renowned research departments or vice versa. Reducing the differences between diverse and evolving institutions to a single score or rank is anathema to many in academia.

Criticisms are largely focused on the methodology of the rankings, the selection of metrics used and the importance ascribed to each of these metrics. Also, the normalization strategy used when creating the ranking can influence the position of a university (Moed 2017). It was argued therefore by Dehon, McCathie, and Verardi (2010) that the position of a university is largely influenced by decisions made by the rankings'

[9] https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2021-methodology
[10] https://www.timeshighereducation.com/content/world-university-rankings
[11] https://www.shanghairanking.com
[12] https://www.topuniversities.com/university-rankings/world-university-rankings/2022

designers. Further, it has been suggested that rankings are biased towards universities in the United States or English-speaking universities, for example, by using a subset of mostly English journals to measure the number of publications and citations (Pusser and Marginson 2013).

In contrast to the above rankings, The Leiden Ranking states that it considers only the *scientific performanc*e of universities and does not take into account other dimensions of university performance, such as teaching performance. Additionally, to appear in the rankings, a university must have produced at least 800 Web of Science indexed publications in the period 2016–2019. Only so-called *core publications* are counted, which are publications in international scientific journals. Also, only research articles and review articles are taken into account. A university needs to have a certain minimum number of scientific publications in order to be included in the Leiden Ranking.[13] It is possible, therefore, that being included in the Leiden Rankings at all is a form of prestige in itself.

However, a central tenet of this work is identifying who gains from OA. To ascertain whether status plays a role in OA, we therefore use the notion of *prestige* to delineate between institutions in terms of status. For this we therefore employ the range of rankings described above as these differ in methodology, allowing for comparisons between them.

2.4. World Bank - Development Indicators

The World Bank Development Indicators are a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The database contains 1,400 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years.[14] The data provides comparable statistics for each country in six areas; poverty and inequality, people, environment, economy, markets and global links. We use GDP per capita from the World Bank Development Indicators for some of the investigations in this study.

2.5. Unpaywall

Unpaywall (UPW) is an open database of links to 29,792,719 open access scholarly articles hosted by more than 50,000 journals and repositories (as of August 2021) and also includes metadata for non-OA articles drawn from CrossRef[15]. Whilst Unpaywall provides links to papers, it does not host any content itself. Unpaywall provides access to this data as a data dump or via an API. We used the snapshot dated to 2021-02-18 for all analyses. Unpaywall assigns an OA Status to every article, which is found in the oa_status field of the API and dataset. There are five possible values: closed, green, gold, hybrid, and bronze. In regards to these terms, Unpaywall states: "*These terms are all commonly used in discussions of open access. Unfortunately, however, there is still no universal agreement on how to define them."*[16] We therefore use Unpaywall's definitions, in the remainder of this report. Additionally, we use the Unpaywall data to check the open access status of papers retrieved via Microsoft Academic Graph using matching based on the DOI of the record. Furthermore, for some analyses we distinguish hosting types: repository only, journal only, or repository & journal hosted content. We classify a publication as being hosted by both a journal and a repository if the "host_type" is "publisher", but the publication also has a repository copy

[13] https://www.leidenranking.com/information/general
[14] https://data.worldbank.org
[15] https://unpaywall.org
[16] https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-

("has_repository_copy" is true). We further classify the remaining publications in either repository only and journal only, depending on whether the "host_type" is "publisher" or "repository"[17].

A known limitation of data from Unpaywall is its dynamic nature - the OA status and type of hosting might change from one data dump to the next. In April 2021, Semantic Scholar removed a substantial number of articles that they had been hosting, resulting in a change from "green OA" to "closed" for about one million articles within the Unpaywall database[18]. This recent drop does not affect our analyses, since the dump we used originated prior to this change. Nevertheless, comparisons of OA shares and dynamics between different versions of the Unpaywall data have to take into account these methodological constraints.

2.6. Responsible Research and Innovation - the MoRRI Dataset

The European Commission defines RRI as "*a process where all societal actors (researchers, citizens, policy makers, business, third sector organisations etc) work together during the whole R&I process in order to better align R&I outcomes to the values, needs and expectations of European society*".[19] RRI operationalises the areas of engagement, gender, science education, open access, ethics and governance.

In 2014, the European Commission commissioned a study on 'Monitoring the Evolution and Benefits of Responsible Research and Innovation' (MoRRI). This evolved in 2020 to SuperMoRRI which, using a range of qualitative and quantitative methods, collected data from research institutes in 28 European countries. These data are the basis for 36 diverse indicators across the areas of focus; engagement, gender, education, open access, ethics and governance. They include straightforward measures such as the share of female scientific paper authorship and citation scores for open access publications, as well as qualitative indicators of public involvement, research ethics and governance mechanisms collected by national experts (European Commission and Directorate-General for Research and Innovation 2012).

The full list of indicators defined by the MoRRI project can be seen in Table 1. The categories are divided into six focus areas: gender equality (GE), science literacy and education (SLES), ethics (E), public engagement (PE), open access (OA) and governance (GOV). Together, these indicators aim to measure the current status quo in RRI across the whole of Europe.

[17] https://github.com/on-merrit/ON-MERRIT/blob/a3c4e2d1cba6ac64e39a2edbdc9a061e8a1e0d62/WP3/Task3.2/spark/jobs/collect_paper_data/__init__.py#L73

[18] https://groups.google.com/g/unpaywall/c/rDOoIgGMxuk

[19] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

Table 1: Breakdown of all SuperMoRRI indicators

	Description of Indicator
GE1	Share of RPOs with gender equality plans
GE2	Share of female researchers by sector
GE3	Share of RFOs promoting gender content in research
GE4	Dissimilarity index (secondary data) – two sub-dimensions
GE5	Share of RPOs with policies to promote gender in research content
GE6	Glass ceiling index (secondary data)
GE7	Gender pay gap (secondary data)
GE8	Share of female heads of RPOs
GE9	Share of gender- balanced recruitment committees at RPOs
GE10	Number and share of female inventors and authors
SLSE1	Importance of societal aspects of science in science curricula for 15–18- year- old students
SLSE2	RRI-related training at higher education institutions
SLSE3	Science communication culture (secondary data)
SLSE4	Citizen science activities in RPOs
E1	Ethics at the level of RPOs
E2	National ethics committees index
E3	RFO ethics index – two sub-dimensions
PE1	Models of public involvement in science and technology decision making
PE2	Policy-oriented engagement with science (secondary data)
PE3	Citizen preferences for active participation in science and technology decision making (secondary data)
PE4	Active information search about controversial technology
PE5	Public engagement performance mechanisms at the level of research institutions
PE6	Not enough data collected and therefore indicator not used.

PE7	Embedment of public engagement activities in the funding structure of key public agencies
PE8	Public engagement elements as evaluative criteria in research proposal evaluations
PE9	Research and innovation democratization index
PE10	National infrastructure for involvement of citizens and societal actors in research and innovation
OA1	Open access literature – two sub-dimensions
OA2	Indicator not used in MoRRI as insufficient data collected - therefore not used for this study
OA3	Social media outreach / take- up of open access literature
OA4	Public perception of open access (secondary data)
OA5	Funder mandates (secondary data)
OA6	RPO support structures for researchers
GOV1	Composite indicator: governance for responsible research and innovation
GOV2	Existence of formal governance structures for RRI within RPOs and RPOs
GOV3	Share of RFOs and RPOs promoting RRI

Having defined these common data sources, we combine and analyse these with a number of different statistical methods to answer the research questions previously introduced. We next move on to present the first study which explores the current OA landscape in terms of the production and citation of OA research papers. We then also examine how the adoption and usage of OA may be influenced by a range of factors such as institutional prestige and geographic location and economic status

3. Who produces and who consumes Open Access?

3.1. Introduction

This study examines the differences in production and consumption of Open Access (OA) literature across geographical boundaries and institutional prestige variables. Our work is motivated by trying to identify whether barriers related to accessing research literature, such as being located at an institution with limited access to non-OA literature, affect the citation behaviour of scholars. Do scholars located in less developed or at less prestigious institutions rely more on OA because their access to subscription literature is limited? Do those who benefit from OA also produce more OA or are production and consumption independent?

We approach our initial study with the use of two paradigms, *production* and *consumption.* In this approach we define production as the publication of OA literature (as a proportion of all research literature produced) by an entity (author, institution, country, continent). We define consumption as evidence of using OA literature by an entity as measured by citation to OA literature.

Using the production and consumption framework, it is possible to measure production and consumption in multiple ways. In our work, we will focus on measuring the *OA Production Rate*, i.e. the proportion of papers produced as OA by an entity. While *OA Production* is somewhat straightforward to measure, there are multiple ways in which *OA Consumption* could be measured. For instance, one option would be to measure the proportion of OA paper downloads by an entity. However, such data are not currently publicly available. As a result, we estimate the *OA Consumption Rate* as the proportion of OA references an entity (authors, institution, country, etc.) used in the research papers this entity produced (as a proportion of total references). The subsequent analyses of OA consumption take as a basic assumption that one can only cite what one has read. We understand this is a somewhat imperfect assumption due to two potential confounding factors: (1) that people may indeed often cite articles that they have in fact not read, and (2) “shadow library” websites (most prominently Sci-Hub). As for the former point, we acknowledge that it has been shown that authors sometimes cite research that they have not read (Ball 2002; Bornmann and Daniel 2008). Given our quantitative methods, we are unable to take account of this factor here. We hence treat this as a limitation of our study. Considering the potential effect of shadow libraries, although our study does not primarily focus on measuring the extent of illegal ways of accessing research literature, which are reportedly common across academia (Nicholas et al. 2019), we take this into account in our data analysis and interpretation. More specifically, the rise of Sci-Hub, a piracy research site offering free access to research content irrespective of copyright, since its creation in 2011 is factored into our study by comparing the pre- and post-Sci-Hub era. Did Sci-Hub have an effect on the citation behaviour of scholars in less developed and at less prestigious institutions? Does Sci-Hub act as a facilitator of illegal access to research literature or does it primarily provide extra convenience in doing so? These are the questions our study will help us to better understand.

3.2. Background

Recent work by Huang et al. (2020) investigated the production of OA literature around the globe based on institutions present in the THE rankings (see section 2.3). They found that, in 2017, the 100 top-ranked

universities made 80–90% of their research publications OA. This figure is slightly higher than the result when looking at the data from the Leiden Dataset which suggests this figure is around 70%.[20]

In 2017, Frenken, Heimeriks and Hoekman (2017) undertook a comparison of institutional performance using data from the Leiden Ranking and found that research performance differences among universities mainly stem from size, disciplinary orientation and country location. The authors state that this result underlines, yet again, that larger universities systematically over-perform in citation rankings. However, the exact cause remains under-researched (Bornmann, Mutz, and Daniel 2013).

Regarding citation behaviours, most studies conclude that OA articles receive more citations than articles that are behind paywalls (Holmberg et al. 2020; Harnad and Brody 2004; Hajjem, Harnad, and Gingras 2005; Kousha and Abdoli 2010). Interestingly however, a *citation disadvantage* for articles in OA journals was found in all research disciplines outside of physics and astronomy (Archambault et al. 2014) and also discovered by van Leeuwen, Tatum, and Wouters (2015), Robinson-Garciá, Jiménez-Contreras, and Torres-Salinas (2016) and Dorta-González, González-Betancor, and Dorta-González (2017).

3.3. Methodology

In this section we investigate the level of production and consumption of OA literature at an institutional, country and continental level. For this study, we employ the research paper dataset from MAG combined with the OA status data from Unpaywall. We use economic data from the World Bank and ranking indicators from THE, Leiden and MAG. We first examine the levels of OA production and consumption (measured as a proportion of all production). We then correlate this at the country and institutional level, and also measure this using THE ranking data. Finally, we examine how GDP per capita is related to the production and consumption of OA literature.

3.3.1. Terminology Used

Table 2: Terminology used

Terminology	Description
Production	The publication of research papers by an entity (continent, country, institute)
OA Production	Research papers produced by an entity. We use Unpaywall[22] data to distinguish between OA and non-OA literature.
OA Production Rate	The proportion of OA research papers produced by an entity.
Consumption	The use of a research paper by an entity as measured by citations in that entity's publications.
OA Consumption	The use of an OA research paper as measured by citations in that entity's publications.
OA Consumption Rate	The proportion of OA research papers used by an entity. In our work, we use as evidence of use the act of *citing* OA research literature in manuscripts produced by an entity. We use Unpaywall[22] data to distinguish between OA and non-OA literature.

[20] https://www.leidenranking.com/ranking/2021/list

We used MAG data to collate all papers with complete metadata to the publishing institution. As previously discussed, this methodology identifies 219m paper / institute pairs, representing 84% of the total MAG corpus. We then used the latitude and longitude from the MAG affiliation metadata to geocode the location of each institution to a three letter country code. This enabled us to group together all institutions from a single country. We aggregate the overall institutional figures to produce an overall mean score for each country.

3.3.2. Ranking Data Used

We use the university ranking data from THE, the Leiden Ranking and MAG as data sources from which we derive our performance / prestige metrics when undertaking comparisons of individual institutions. We also use GDP data from the World Bank[21] in order to segment the results allowing comparisons between countries and continents. Data regarding institutions, authors and articles for these experiments come from the MAG dataset (Sinha et al. 2015b) which as of June 2021 contains 260,423,032 papers. The coverage of MAG data has been shown by (Hug, Ochsner, and Brändle (2017) to be comprehensive, and comparable to both Scopus[22] and Web of Science (WoS)[23].

It should be noted that Ranjbar-Sahraei, van Eck, and de Jong (2018) found that a number of publications in MAG have missing or incorrect affiliation information when looking at a sample of outputs for a single university. One limitation of this study therefore is based on the fact that, whilst we can remove data with missing or incomplete information, it is not possible to validate the accuracy of every single record. Using MAG data, we collated metadata and all known citations for all papers where the institution and author data were complete. From the complete MAG data, we were able to collate identifiers for 219m papers by 44m authors from 24,000 individual institutions (all figures are close approximations).

Finally, in this next section, we utilise the MAG rankings, in which additional scores are calculated for a range of entities. An entity's *importance* is calculated using its relationships with other entities in the graph. For example, a paper entity recently published in Nature receiving a high number of citations is likely to have high importance, whereas a preprint paper entity not associated with a conference/journal is likely to have a low importance.

Which ranking system and methodology is best overall, or even fit for purpose, is a long-running and far reaching debate that falls outside of the scope of this study. It is, however, useful to use a basket of these indicators as a proxy of 'prestige' which can allow for a closer look at how this prestige is related to both the production and consumption of OA literature.

3.3.3. Selection of the time periods for the following studies

The choice of time periods for the following studies in this study were selected to visualise the potential influence of a Sci-Hub effect after 2011. Sci-Hub enables users to freely download PDF versions of research papers, including primarily those that are closed access and locked up behind paywalls. Sci-Hub was launched in 2011 and has grown rapidly since. Estimates by Himmelstein et al. (Himmelstein et. al, 2017) report that Sci-Hub contains 68.9% of the 81.6 million scholarly articles registered with Crossref and 85.1% of articles published in toll access journals.

[21] https://datahelpdesk.worldbank.org
[22] https://www.scopus.com/
[23] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

There has been significant research on the usage of Sci-Hub. Nicholas et al. (2019) found that Sci-Hub use was increasing and that a quarter of the early-career researchers they interviewed as a part of their study now use it. In terms of the global usage of Sci-Hub, research by Bohannon (2016) found a quarter of the Sci-Hub requests for papers came from the 34 members of the Organization for Economic Co-operation and Development (OECD), some of the richest nations. The usage of Sci-Hub is shown to have global coverage; data from Sci-Hub released in 2016 demonstrated that many of its users are in the U.S. and Canada and these users tend to be located near university campuses, suggesting they may not have institutional access to these articles despite their location. With OA citations rates, we can only assess the research that authors have access to. What we cannot do is to assess the access or lack of it to research literature of non-authors.

We therefore divide the MAG dataset into three time periods; 2006-2010, 2011-2015 and 2016-2020.

3.4. Results

We examined OA Production and OA Consumption across over 190 countries and from six continents using the MAG dataset. For the experiments in sections 3.4.1 and 3.4.2 we focus on the most recent period from 2016-2020.

3.4.1. Total OA Production and OA Consumption - breakdown by country and continent

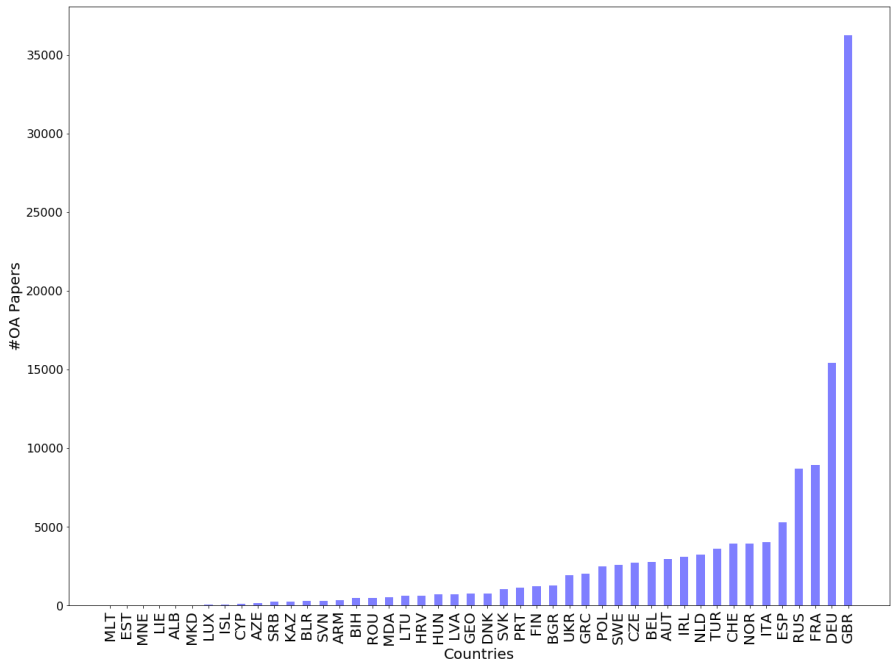

Figure 2: Overall production of OA papers for each country in Europe.

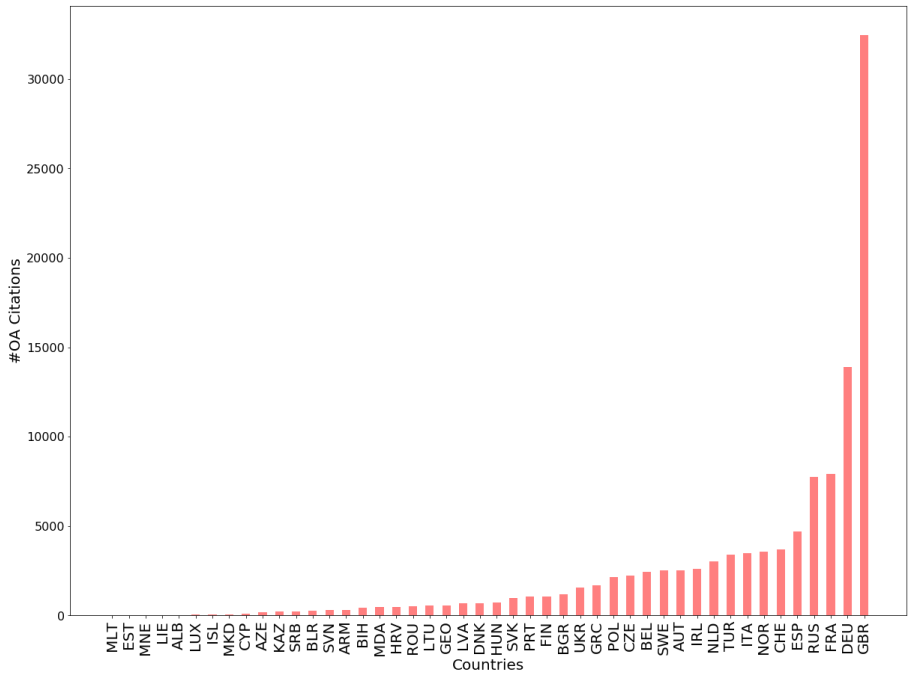

Figure 3: Overall citations to OA papers by each country in Europe.

Figure 2 and Figure 3 show that in terms of absolute numbers, the UK and Germany are the leaders in both the production and consumption of OA research papers. They are closely followed by France and Russia. However, as these are the largest European countries by population, we were motivated to explore also OA Production and OA Consumption rates, i.e. the proportions of OA papers produced and cited by a country.

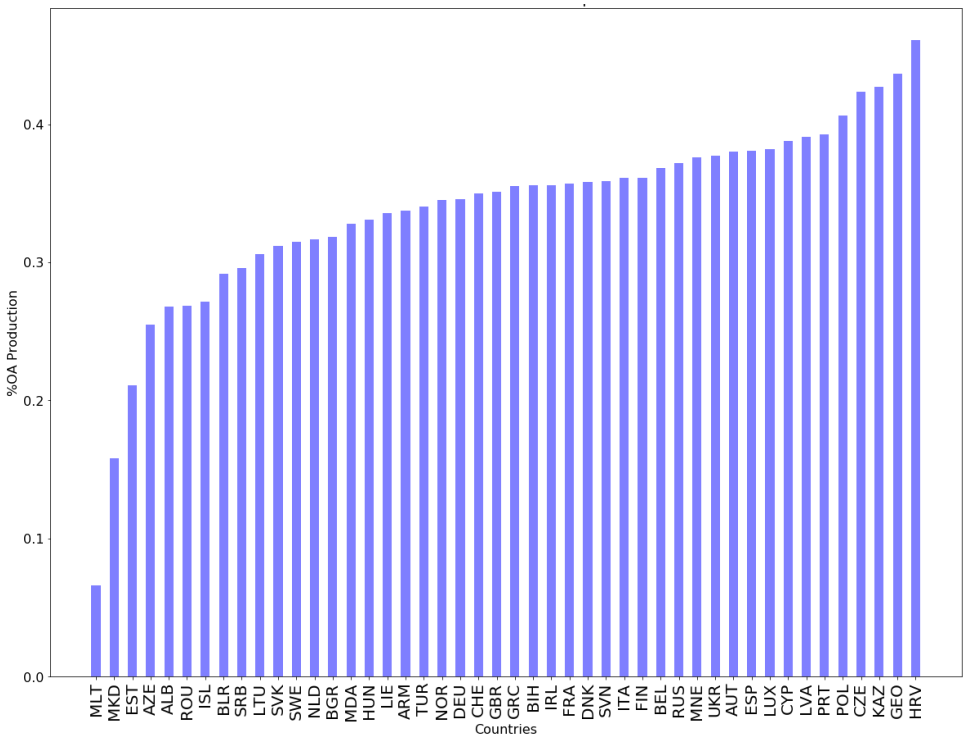

Figure 4: Proportion of OA papers produced per country for Europe.

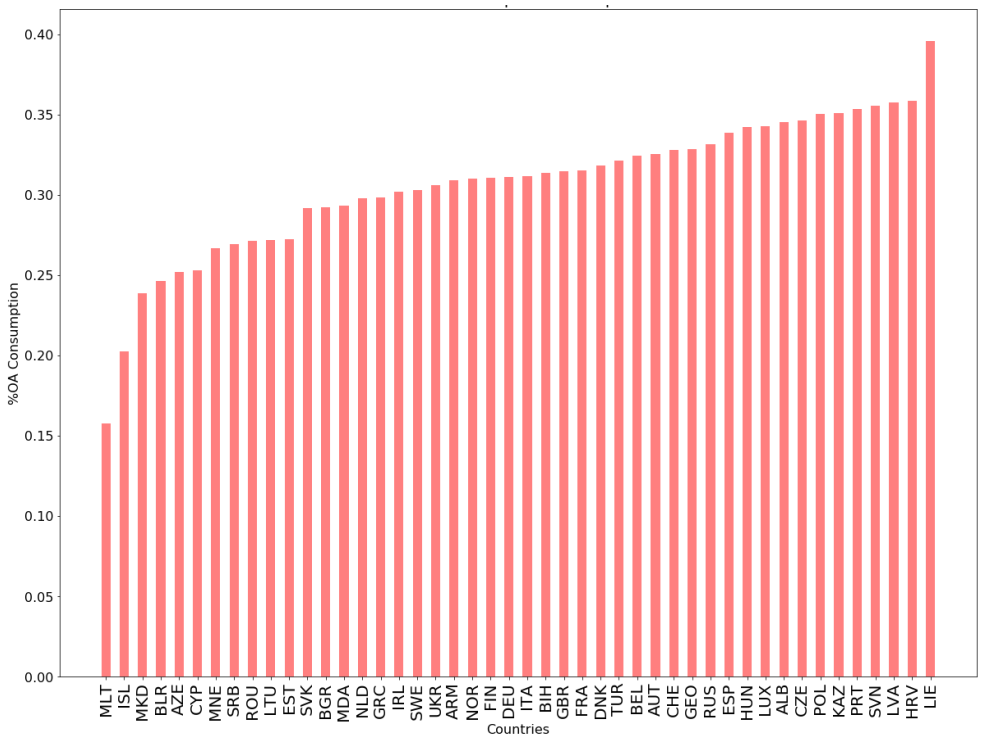

Figure 5: Proportion of OA papers consumed per country for Europe.

If we therefore measure the production and consumption of OA research papers per country as a proportion of the overall production and consumption of research papers per country (Figure 4 and Figure 5) it can be seen that a different story appears. Taking the United Kingdom as an example, only a third (32.3%) of all research papers are actually OA (in the MAG dataset) with many other European countries outputting a higher proportion of OA research papers. Whilst only the results for Europe are shown above for brevity, data for all continents and countries can be found in the Annex.

3.4.2. *net_OA Production* vs. *net_OA Consumption* - breakdown by country and continent

Having established a difference between OA production and consumption rates compared to closed (or non-OA) production rates, we now define *net_OA consumers* as those who are on average more likely to consume OA literature (measured as the proportion of OA papers they cite in their research) than they are to produce OA literature. A *net_OA consumption rate* is calculated thus;

$$oa\_production/total\_production - oa\_consumption/total\_consumption$$

This allows us then to calculate whether a country, continent or institute is therefore a *net_producer*, or *net_consumer* of OA literature for the period from 2016-2020 being measured in this section. In the following figures, the *OA Net Production Rate* is plotted for each country. Countries who are *net producers* of OA research literature appear to the right of the inflection point on the zero line whilst those who are *net consumers* appear to the left. We observe that the majority of countries are in fact net producers of open access scientific literature. This was an unexpected finding.

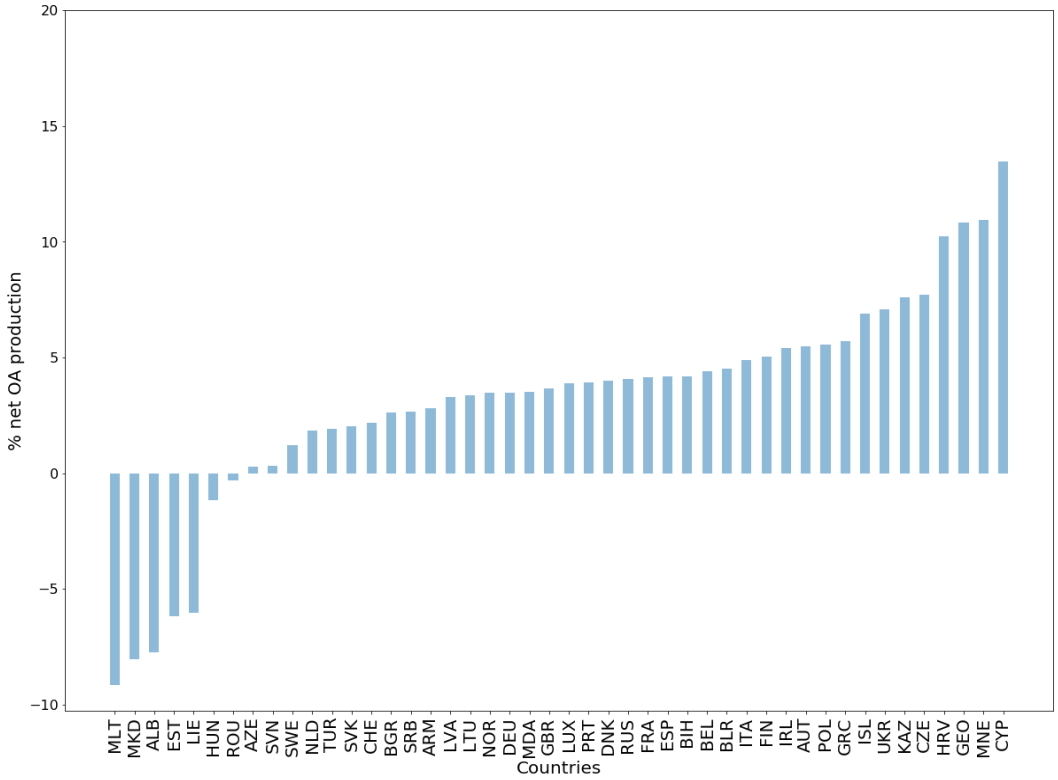

Figure 6: Net OA production for Europe by country

Figure 6 shows the net OA production rate plotted for European countries. It can be seen that only Malta, North Macedonia, Albania, Estonia and Lichtenstein are net consumers of OA research literature. Most European countries are producing between 1%-5% more OA than they are citing. Hungary, Georgia, Montenegro and Cyprus on the right of the graph are producing between 10%-14% more OA than is being cited.

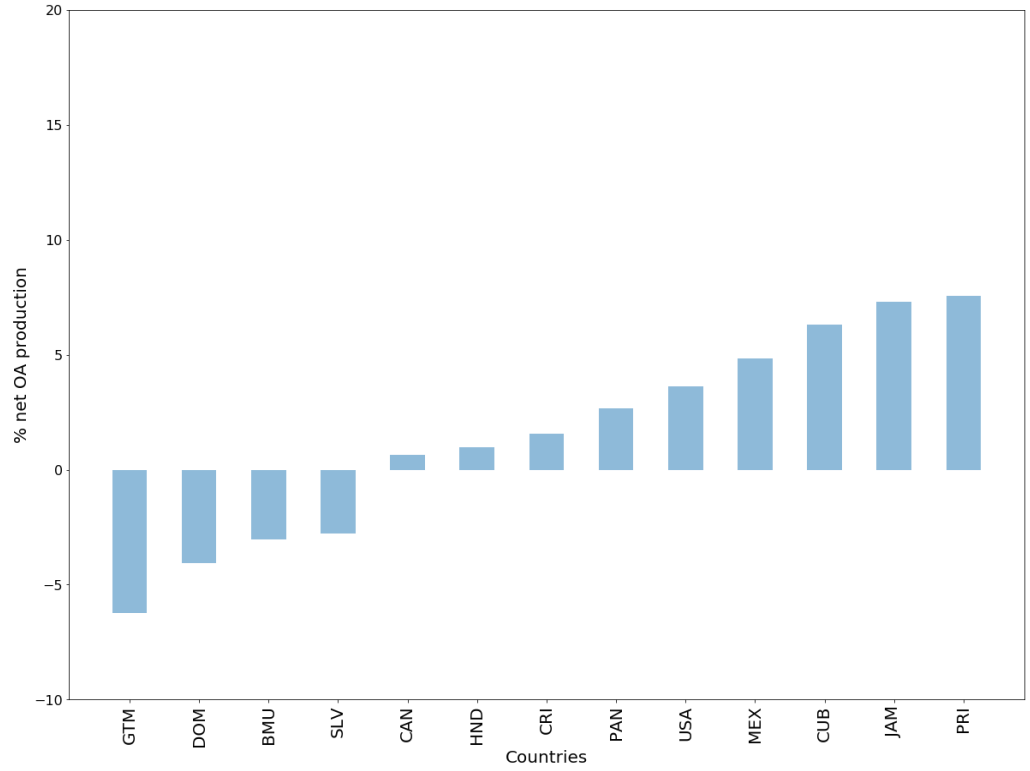

Figure 7: Net OA production for North America by country

We find similar results across other continents too. Both North America (Figure 7), Asia (Figure 8)and Africa (Figure 9) show that most countries are net producers of OA research literature.

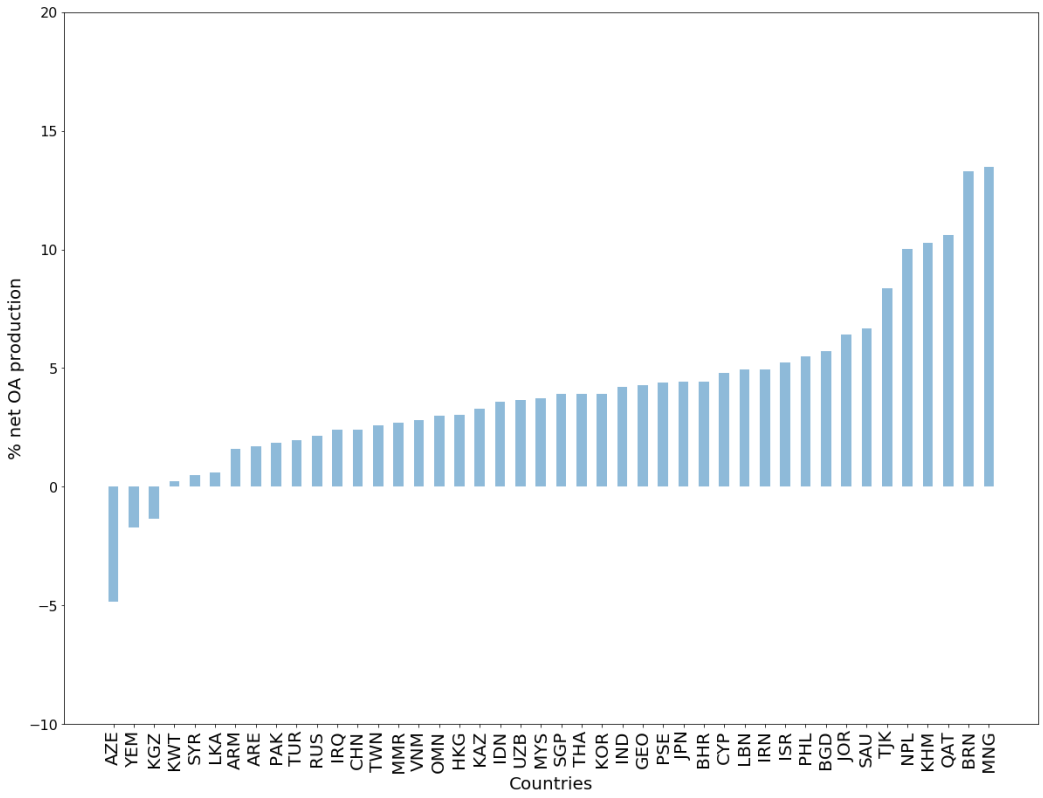

Figure 8: Net OA production for Asia by country

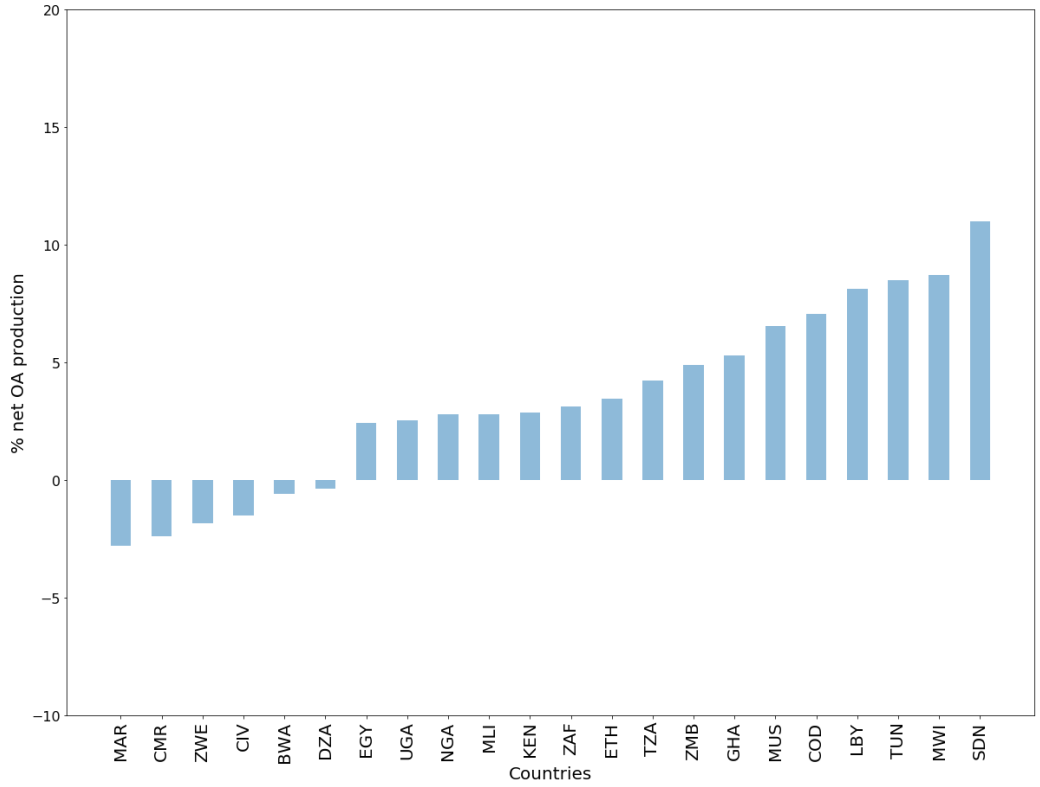

Figure 9: Net OA production for Africa by country

In this next section we now look at the correlation between OA consumption and production by aggregating institutional data from MAG. We then combine this with ranking data to examine any link between the *prestige* of an institution and their OA practices.

3.4.3. Correlation between OA production and consumption.

Next, we examine the correlation between *OA production* and *OA consumption* at a continental level using aggregated data from the institutional level across all three time periods. (Figure 10, Figure 11 and Figure 12). Each individual dot represents a single institution in a country and each colour represents a continent as shown in the legend.

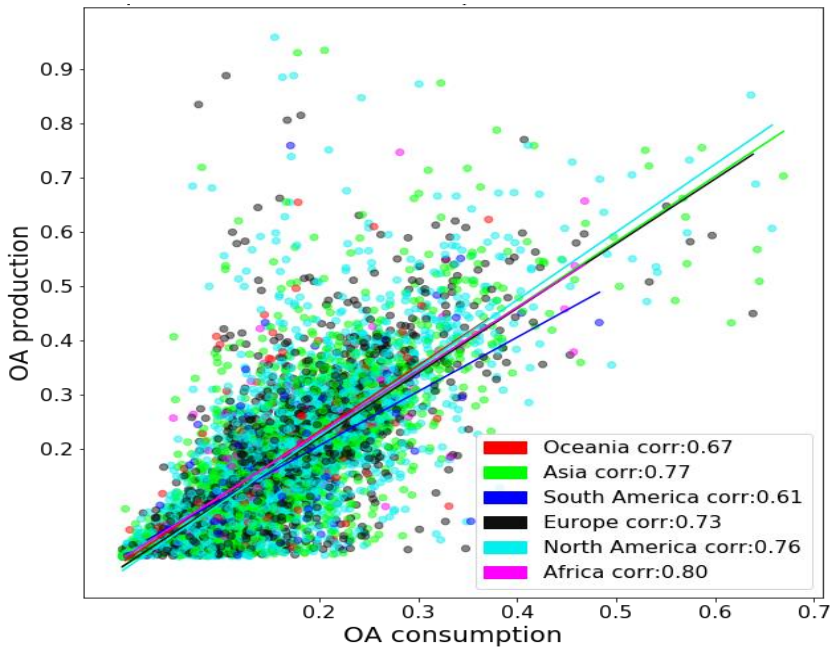

Figure 10: Correlation between OA production and consumption by country and grouped by continent 2006-2010. (n=190, r=0.75)

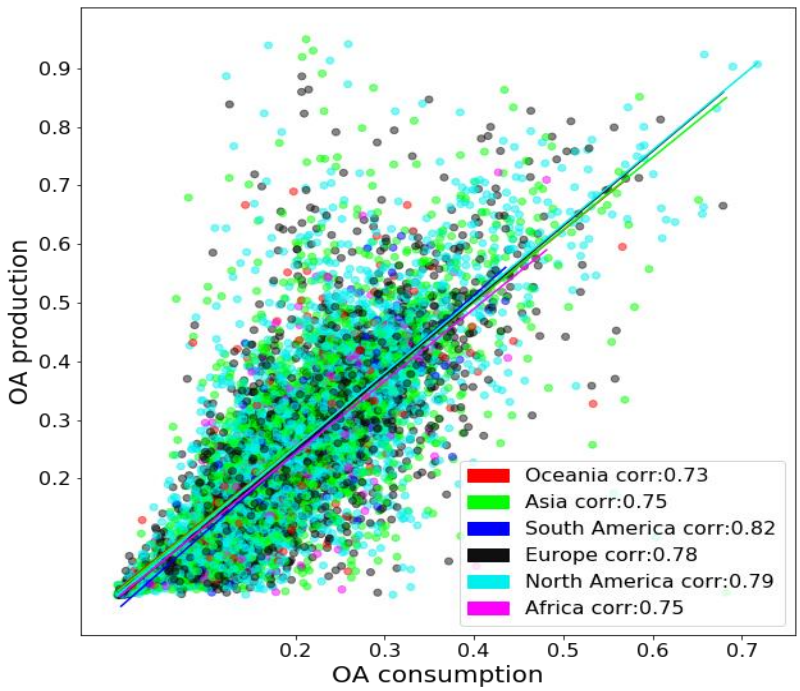

Figure 11: Correlation between OA production and consumption by country and grouped by continent 2011-2015. (n=190, r=0.77)

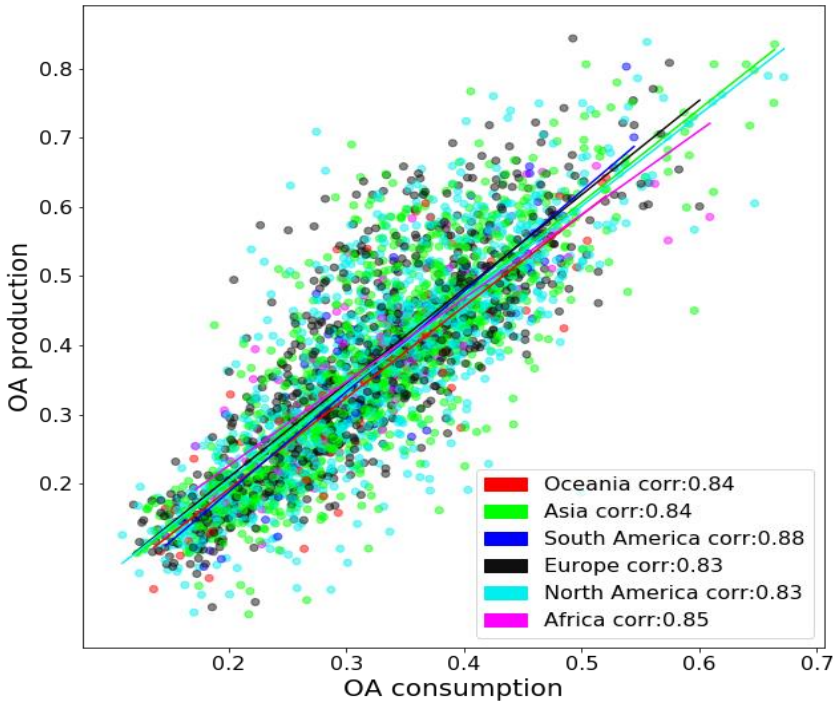

Figure 12: Correlation between OA production and consumption by country and grouped by continent 2016-2020. (n=190, r=0.84)

We find a statistically significant correlation between the rates of OA production at the institutional level and the rates at which OA literature is then cited by this institution. Further, the strength of this correlation has increased markedly in recent years; for the 190 countries covered by our study this correlation increased from *r=0.77* to *r=0.84, n=190*. This suggests that institutions that are producing more OA literature tend to also use more OA. We find only small differences in the correlation rates for different continents and again this gap has continued to close in recent years.

3.4.4. Correlation between institutional prestige and OA consumption.

Having identified a correlation between the production and consumption of OA research literature, our next investigation was to determine whether there was a link between the prestige of an institution, using a range of different ranking methodologies, and the levels of OA consumption at these institutes. The timescales for this investigation maintained the same segmentation as in the previous section.

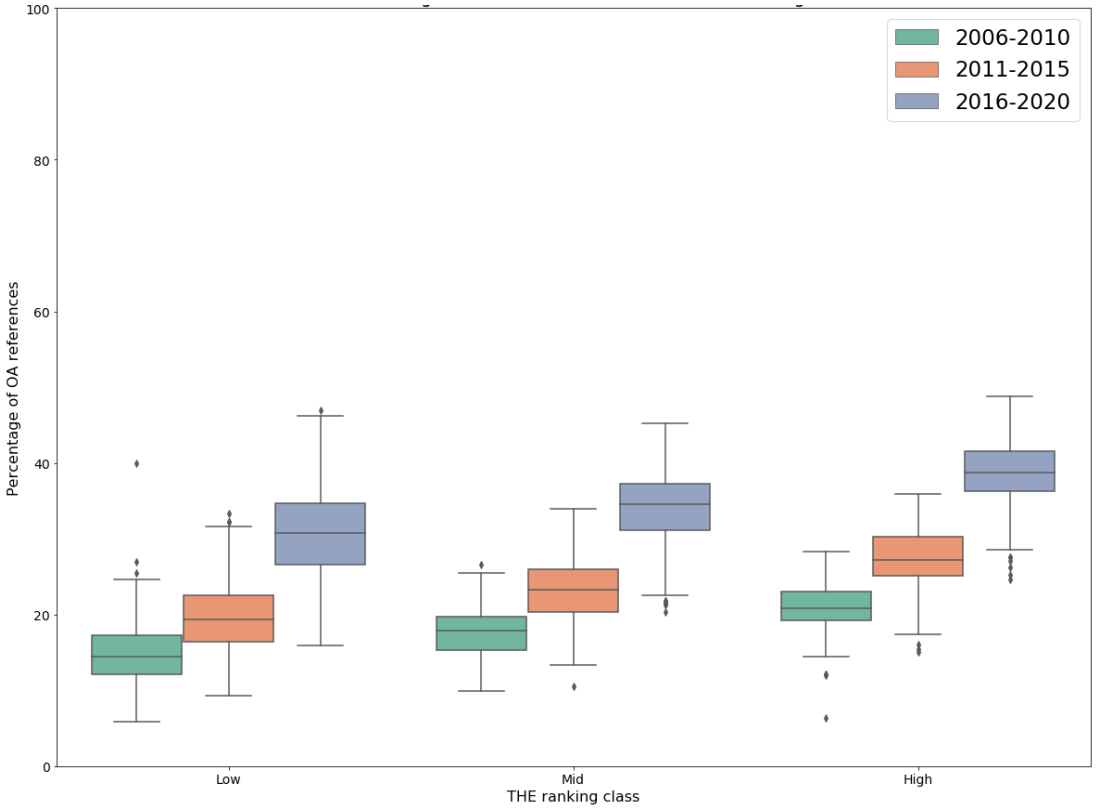

Figure 13: Box plot for OA consumption for institutions based on THE ranking

When using ranking data from the Times Higher WUR, we find a statistically significant difference in the amount of OA content cited by differently ranked institutions. Institutions ranked in the top third on average cite 13% more open access content than those in the bottom third. (Figure 13)

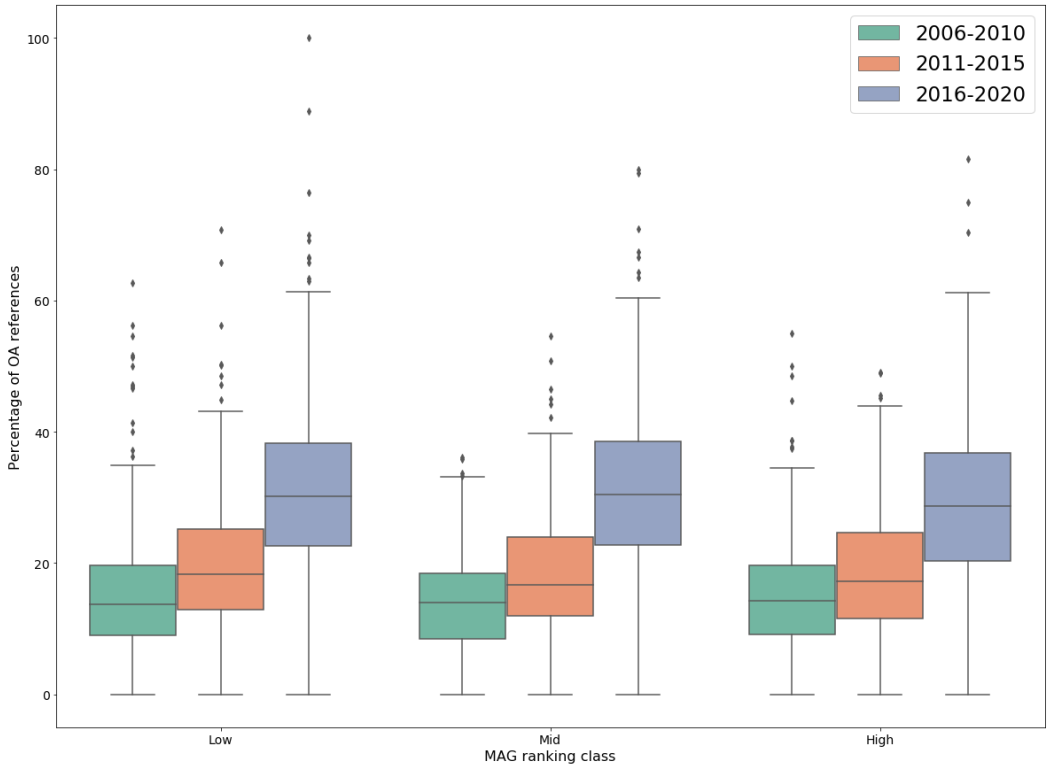

Figure 14: Box plot for OA consumption for institutions based on MAG ranking

Figure 14 uses the same dataset of papers, authors, institutions and citations but uses the ranking taken from MAG. Figure 15is created based on the same dataset and ranking data from the Leiden Rankings to calculate the rank of each institution.

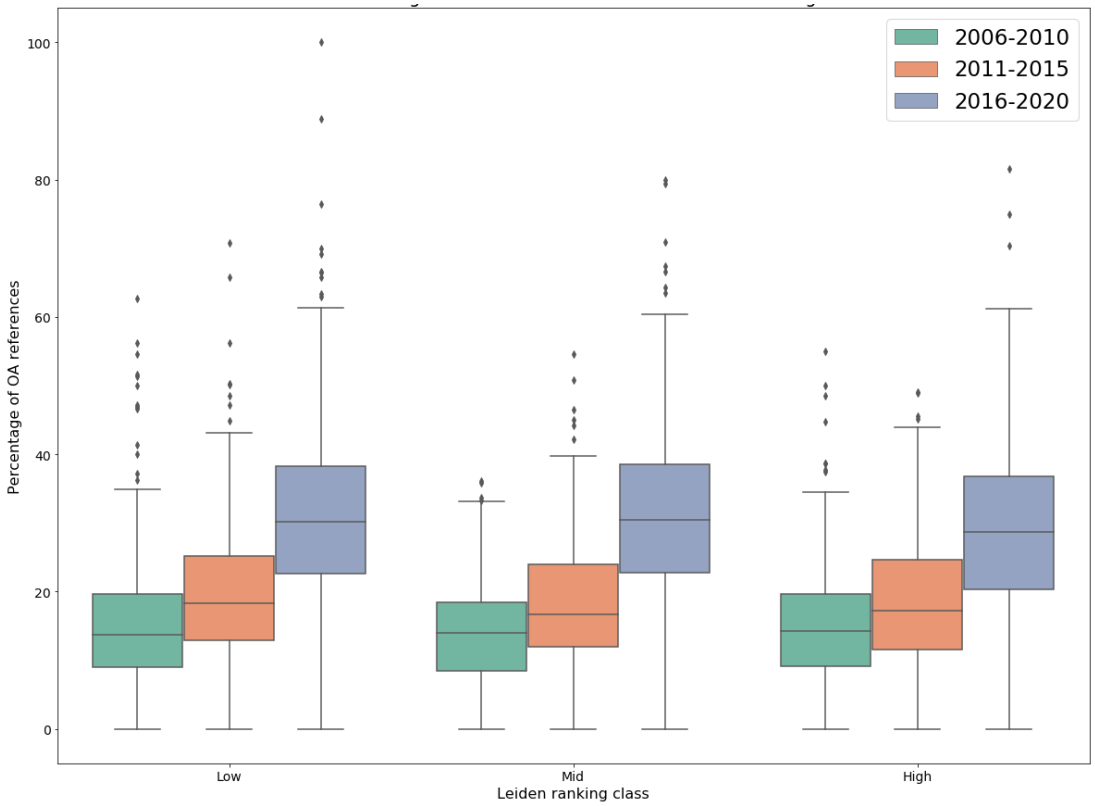

Figure 15: Box plot for OA consumption for institutions based on Leiden ranking

The Leiden Ranking and the ranking data extracted from MAG both produce a ranking based on bibliometric data. The THE WUR uses a proprietary ranking system, the exact calculations for which are not publicly available. It is therefore an interesting result that we only observe a difference in citation rates when using the THE data. If we only use this result, this would suggest that lower ranked institutes tend to cite a smaller percentage of OA research papers than their higher ranked counterparts which seems counterintuitive. However, it should also be noted that the speed with which OA is growing (growth between periods) is far more significant than the difference in the rankings, and it continues to accelerate.

3.4.5. Correlation of Publication vs Citation rates by institution over time

We next examine the correlations between the production and consumption of OA literature for institutions in the Times Higher WUR rankings using Pearson's *r*. The following figures show the correlation between OA production and OA consumption at the institutional level. Each dot is a single institution and covers all institutes in THE WUR rankings. The dots are coloured and sized according to the institutions' ranking.

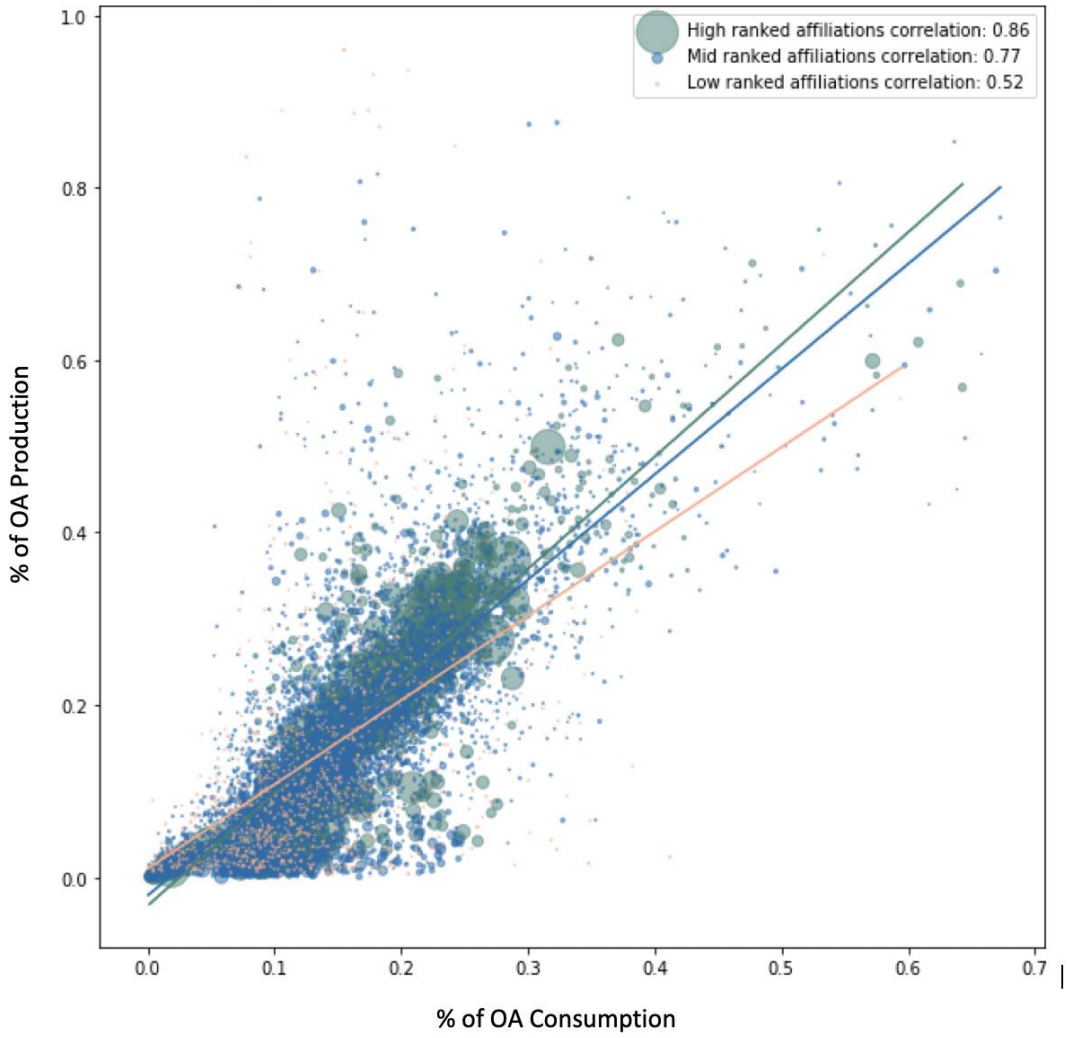

Figure 16: Correlation of OA production vs. OA Consumption 2006-2010 based on THE ranking

It can be seen from Figure 16 that for both periods covered, there is a stronger correlation between production and consumption for the higher ranked institutions (*r=0.86)* than mid (*r=0.77) and* the lower ranked ones (*r=0.52)*. It can be seen from Figure 17 that this gap closed somewhat in the second time period. This change was largely driven by lower ranked institutions increasing rates of production of OA literature. This is borne out in Figure 17 where it can be seen that many more lower-ranked institutions (orange dots) increased the output of OA literature compared to the earlier time period.

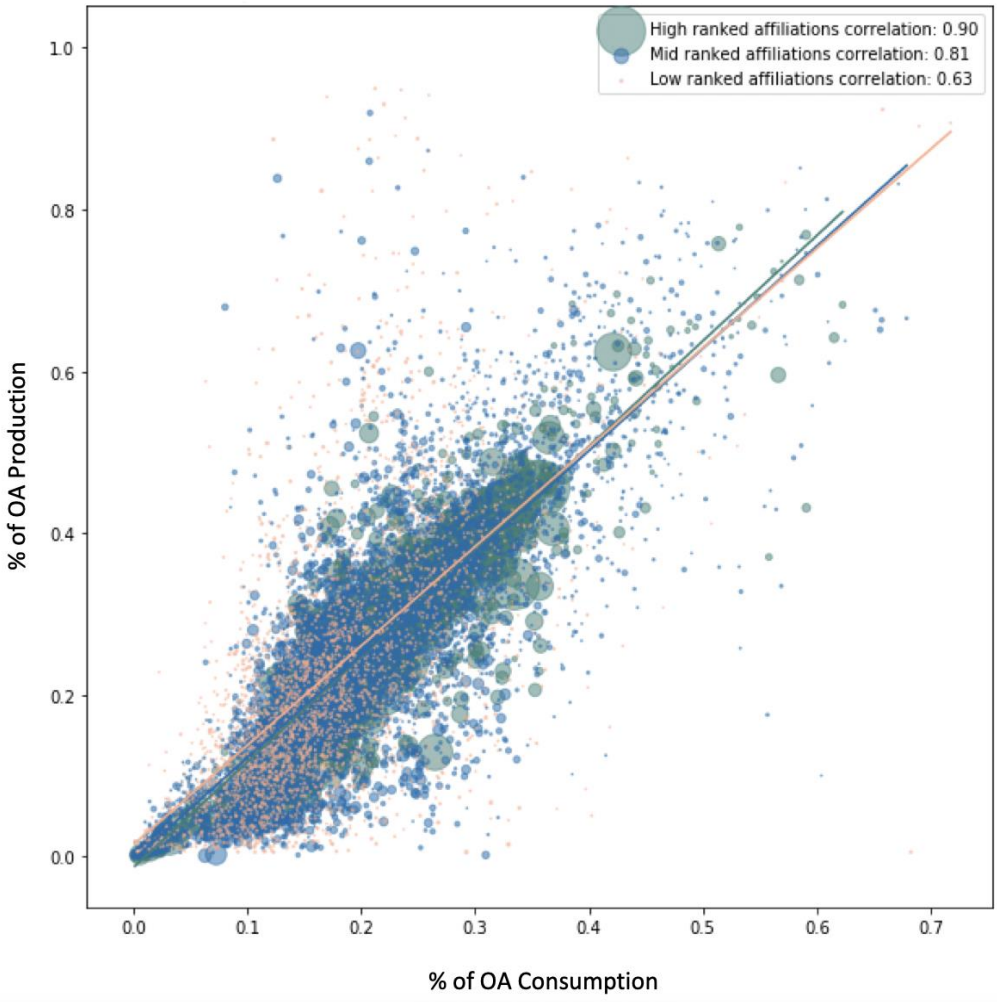

Figure 17: Correlation of OA production vs. OA Consumption 2011-2015 based on THE ranking

Overall, however, the lower ranked institutions both produce and consume less OA when measured using the THE World University Rankings. There are several reasons why this may be the case. Higher ranked institutions were early adopters in building OA infrastructure and consequently realised its benefits earlier than the lower ranked institutions. The size, wealth or location of the institution in question are all potential confounding factors here and these differences remain to be investigated in future work.

3.4.6. Rates of OA citation vs. GDP per capita

The MAG dataset allowed for the collation of data about institutions in a total of 190 countries from all continents, allowing us to ask whether higher rates of OA citation are associated with lower rates of GDP per capita. If that were the case, this would suggest that researchers in countries with low GDP per capita might be disadvantaged in accessing subscription literature and that they might therefore lean towards citing more OA literature instead.

Contrary to our initial intuition, we observe, as shown in Figure 18,Figure 19 and Figure 20, that there is, in fact, no correlation at a country level between the consumption of open access content and GDP per capita,

based on World Bank data[24]. The data for these charts are based on the GDP per capita figures for 2020 for all countries. There are some distinct outliers such as COG (Congo) and RWA (Rwanda), however, this is likely due to the sample size, i.e. there are not enough articles for an individual country to make a valid calculation.

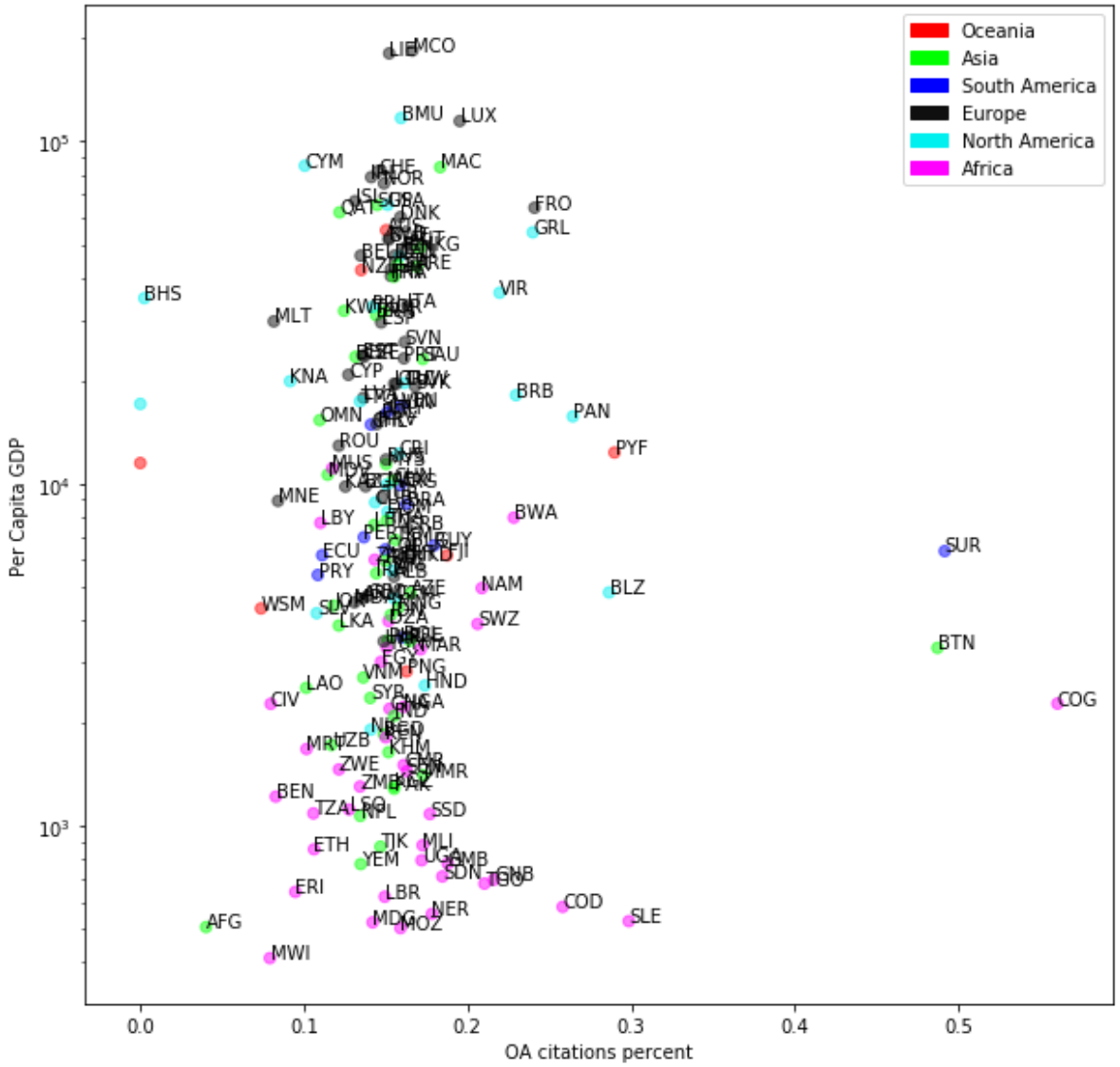

Figure 18: OA consumption vs GDP per capita 2006-2010

[24] https://data.worldbank.org

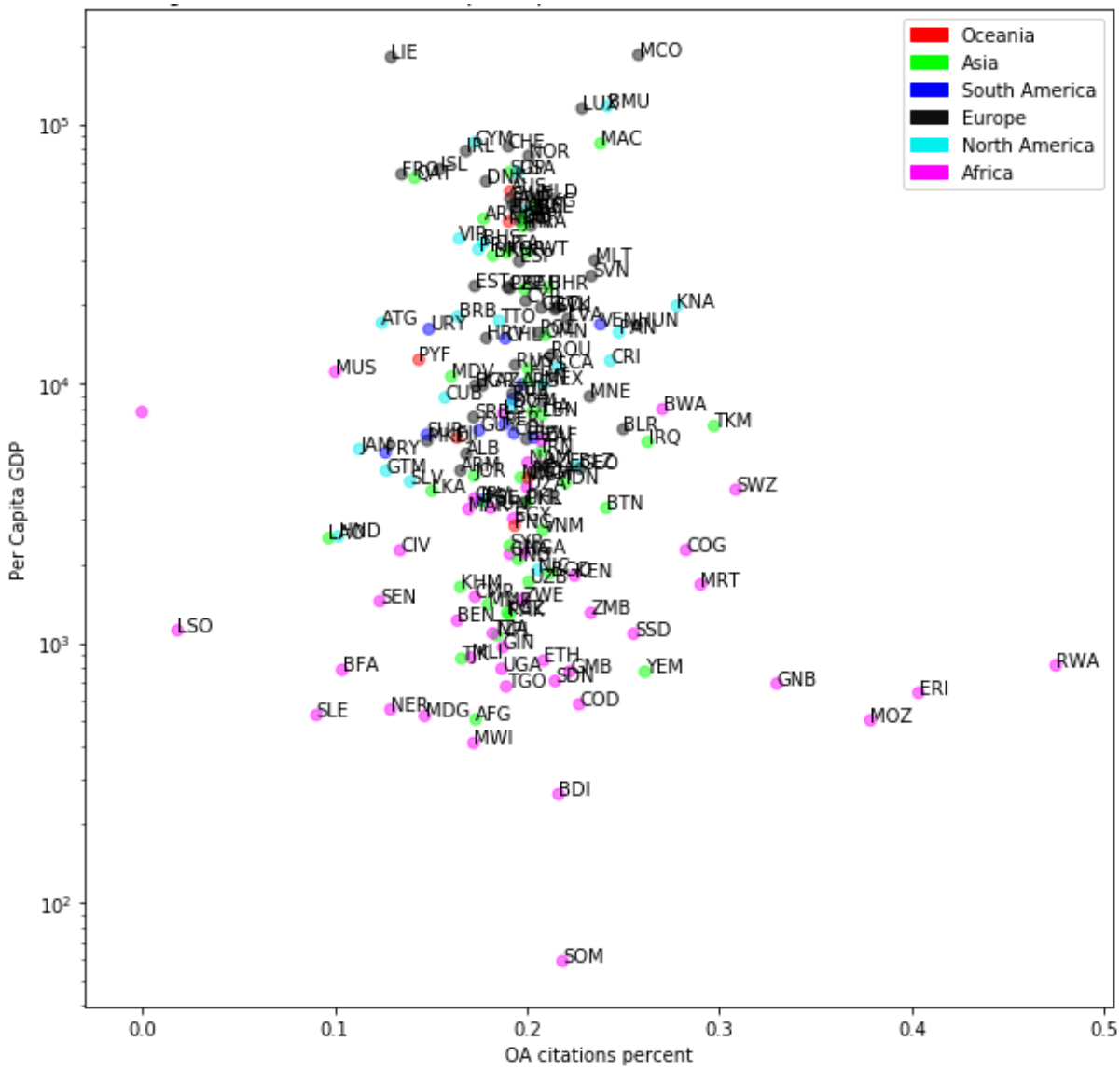

Figure 19: OA consumption vs GDP per capita 2006-2010

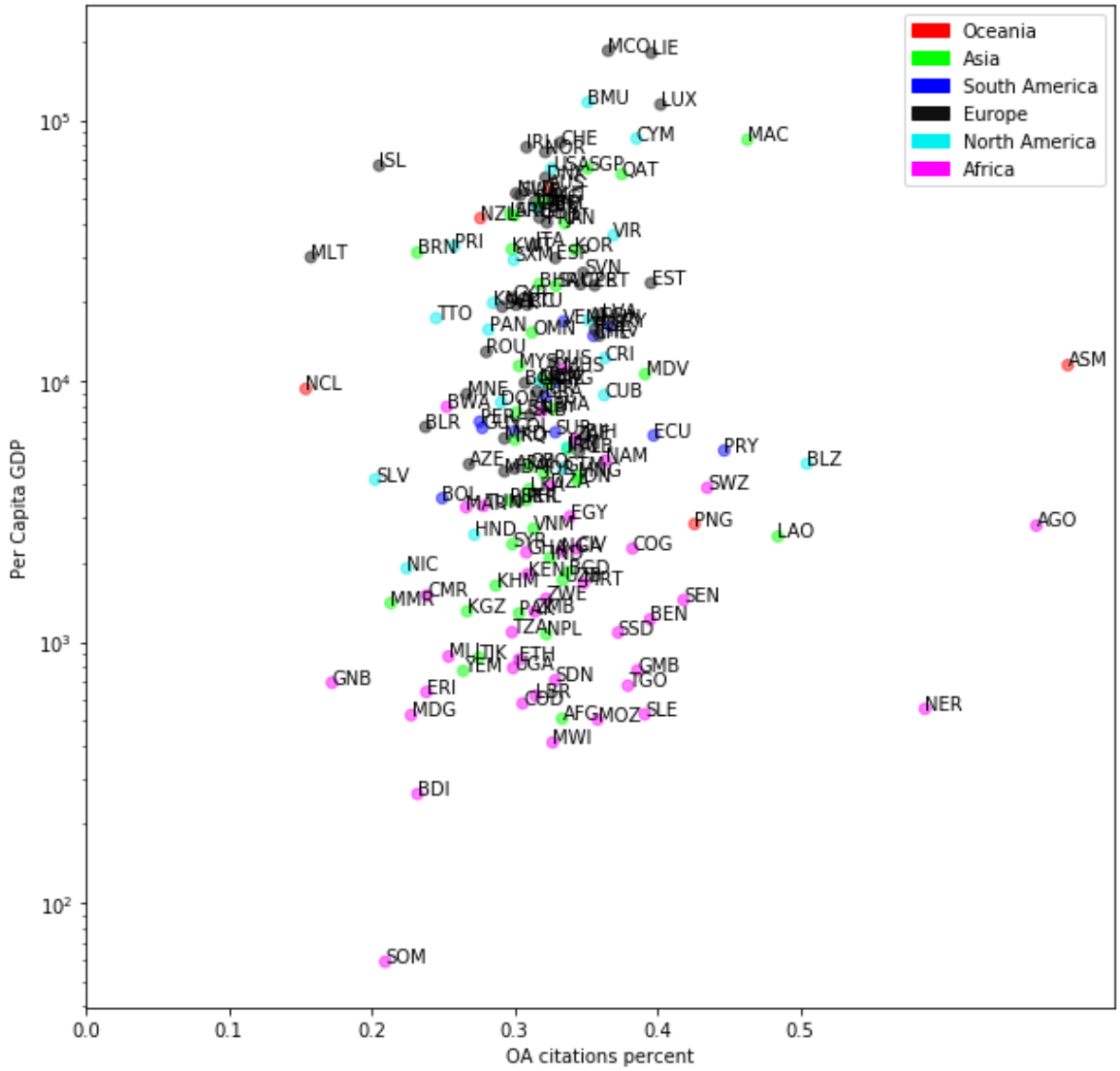

Figure 20: OA consumption vs GDP per capita 2015-2020

It can be seen that, overall, we find no correlation between GDP per capita and the rate of OA consumption. As shown in the previous section, there is a medium-strong correlation between production and consumption at the institutional level. This suggests that there are higher performing institutions in countries with lower economic status and vice versa.

3.5. Discussion

We first examined the overall rates of the production and consumption of OA literature for 190 countries across six continents. In the preceding sections we have demonstrated that, despite the continued growth of OA, most institutes and countries remain to be net OA producers. For some of the largest producers of OA research papers, notably the UK and Germany, the proportion of OA research output as a total of all research output is lower than expected. Piwowar et al. (2018) found the proportion of OA literature to be around 45%, this is higher than our findings however their study used data from Web of Science (WoS) and CrossRef.

Importantly, we find a strong correlation between the OA Production and OA Consumption rates. This correlation appears stronger for higher ranked institutes, particularly when one considers those institutes contained within the Times Higher Rankings.

Using the previously described frameworks of production and consumption, our original hypothesis was that high-income countries would be net producers of OA and low and low-middle income countries would be net consumers. This hypothesis was motivated by our initial assumption that high income countries have more resources to pay for APCs and to make the investment into OA infrastructures. Further, our expectation was that low and low-middle income countries face significant barriers in accessing subscription-based literature and that it will be possible to observe this in the literature they cite in the manuscripts that they produce. However, we cannot confirm this hypothesis from the analysed data.

The narrative that has formed throughout this initial study is that the production and consumption of OA literature is highly correlated at the institutional level and that this correlation has continued to strengthen in recent years. We observe the more highly ranked institutions, when using THE rankings, are marginally greater producers and greater consumers of OA than lower-ranked institutions. One explanation for this phenomenon might be that higher ranked institutions had the resources to invest in OA, that they became the first movers, advocates and adopters of OA, and that their strategy is being followed by the lower ranked institutions. However, this difference is less significant when Leiden or MAG institutional prestige indicators are used. However, this may also be due to the differences in the way the rankings are calculated, with the MAG and Leiden Rankings being purely based on bibliometrics whereas the THE rankings use a broader range of qualitative and quantitative indicators.

A recent study by Siler et al. (2018) showed that, for the field of Global Health, lower-ranked institutions are more likely to publish in closed outlets. Their rationale here is that this is due to the cost of article processing charges (APCs) levied by the publishers. A detailed breakdown of the effects of different OA statuses and APCs is covered in section 6. We observe only weak correlations between the wealth of a country, using GDP per capita as the measure, and the citation of OA literature. This is a somewhat surprising result indicating that the OA economy is much more strongly divided between prestigious and non-prestigious institutions than between the high-income and low-middle / low-income countries.

- Across all considered rankings in all categories low, mid, high tier, there has been an increase in OA consumption over time.
- Based on the graphs in section 3.4.1, we can observe that higher ranked institutions, according to the THE rankings, tend to be marginally consuming more OA (and therefore benefiting more) than the lower ranked institutions at any given time. This gap is smaller, but still present when alternative rankings such as Leiden or MAG are used.
- Remarkably, while we found strong correlations between an institutional rank and OA consumption, contrary to our intuition, we find no correlation between GDP-per capita and OA consumption. This shows that OA adoption is not (or no longer) divided on the rich country vs. poor country perspective, but rather on the prestigious vs. not-prestigious institutional axis.
- Analysing pre-Sci-Hub and post-Sci-Hub OA consumption rates, contrary to our intuition, we observed no “Sci-Hub effect” at a country level. More specifically, the OA-consumption rates for lower GDP per capita countries as well as lower-rank institutions followed a similar trend to the higher GDP per capita countries as well as higher-rank institutions in the pre-Sci-Hub and post-Sci-Hub periods. This suggests that while Sci-Hub might have simplified the process of accessing scientific literature for those who cannot afford subscriptions, it is possible that other, perhaps more informal, sharing networks existed even prior to Sci-Hub playing a similar role. From this perspective, it is

possible that Sci-Hub acted more as a facilitator simplifying the process of accessing subscription literature for those without access rights, rather than making a material change for authors of research papers. However, this finding might not apply to those outside of academia who might not be that well connected as academics affiliated with research institutions.

Publication and citation behaviour is clearly influenced by a wide range of factors. In this initial section we considered all publications by all institutions. As demonstrated by previous studies, the research domain may be of considerable influence, with bio-medicine and the health sciences in particular exhibiting differing patterns of OA production and consumption behaviour. We explore some of these specific areas in greater depth in Section 6.

4. How is institutional performance related to the application of RRI policies and OA publishing?

4.1. Introduction

In the following section, we conduct an exploratory analysis combining European RRI indicators from the MoRRI dataset with ranking data from the Leiden Ranking. More specifically, we focus on how traditional performance and prestige indicators, as recorded in the Leiden Ranking, are associated with OS and RRI factors recorded in the MoRRI dataset. The work is motivated by the question of how the application of OS and RRI practices is associated with performance and prestige. We consider a range of OS and RRI indicators including gender, academic and industry collaboration, policy and societal engagement. Our investigation is carried out at the granularity of institutions and countries, reflecting the granularity at which RRI data have been captured.

We find strong correlation between several RRI pillars, such as *public engagemen*t and *RRI policies* and *institutional prestige* and *OA production.* Surprisingly, we do not observe a clear association in the data between how a country scores on gender equality policies and its actual female/male diversity. However, this might be due to these policies being still relatively new to have a sufficiently profound impact on the composition of the research workforce.

4.2. Background

In 1994, in the context of the 4th EU Framework Programme, the 'ELSA' programme was introduced as a label for developing and funding research into the Ethical, Legal and Social Aspects of science and technology. The ELSA label has been adopted by other funding initiatives as well, notably in European partner countries. Although the acronym ELSA was coined by funding agencies rather than by the research communities, it nonetheless managed to evolve into a recognisable approach (Zwart and Nelis 2009). As noted by Zwart, Landeweerd, and van Rooij (2014) in 'Adapt or perish? Assessing the recent shift in the European research funding arena from 'ELSA' to 'RRI', rather than being an 'empty signifier', ELSA actually came to signify something, namely a particular research practice. ELSA evolved over the first decade of the 21st century and gained some traction amongst the research community. From 2012, a new push towards "Responsible Research & Innovation" (RRI) gained pace. In terms of defining RRI, it is not a new discipline or field but a strategy to be adopted to drive change in the way research is funded and assessed. There is a clear overlap between the remit of the two programs however RRI is focused on the area of societal responsibility. Under the umbrella of RRI, the European Commission identifies gender equality, open access, ethics, science education and public engagement as areas for focus. The EU MoRRI (Monitoring the evolution and benefits of RRI) project ran from 2014-2018, aiming to identify benefits associated with the implementation of RRI practices. This was followed in 2020 by the SuperMoRRI project, which first defined an improved set of RRI indicators, then developed and implemented a system to collect quantitative and qualitative data from different levels across European countries. It is this data that we use for the investigation in the following section.

It is with this in mind that we investigate the connection between current RRI indicators, open access science and performance or prestige at an institution, in particular how academic performance (evidenced using a basket of indicators) is associated with application of RRI and Open Science principles. We then look at the application of RRI and Open Science principles along criteria of geographical location, gender, institutional standing for a range of European countries and find significant differences in the adoption of RRI practices

4.3. Methodology

We conducted an exploratory analysis looking at how RRI indicators are associated with prestige indicators at the institution level. Each of the SuperMoRRI indicators is compiled from a range of quantitative and qualitative data. These data are then aggregated into a score for each country for each area; gender, impact on society, public engagement, ethics, open access and government RRI policies. We then combine these with data from the Leiden Ranking covering 358 institutions from 24 European countries. As described earlier, there is a lower bound for an institution's inclusion in Leiden Ranking in regards to the quantity of publications required, namely an institution must have at least 800 Web of Science indexed publications from 2016-2019. A clear-cut result of this approach can be seen for the UK where only 58 of 157 UK HEIs are included in the rankings. One limitation is that while Leiden data is available at the institution level within each country, the SuperMoRRI dataset is aggregated to the country level. Additionally, the RRI policy data was made available in 2018 and offers a single 'snapshot', while the Leiden Rankings have a temporal facet.

Table 3: Indicators for this section of the Study, the source of this data and a brief description of each.

Indicator	Source	Description
F_gender	SuperMoRRI	Combined metric of 10 indicators describing the quality of gender policies in a country.
F_society	SuperMoRRI	Combined metric of 4 indicators describing the quality and impact of science in society.
F_engagement	SuperMoRRI	Combined metric of 9 indicators describing the quality of public engagement in a country.
F_ethics	SuperMoRRI	Combined metric of 5 indicators describing the use and quality of ethics in each country.
F_policy	SuperMoRRI	Combined metric of 3 indicators describing the use of RRI at the organisational and governmental level.
F_prestige	Leiden ranking	Normalised number of publications belonging to the top 10 publishing venues across the institutions of each country.
F_collab	Leiden ranking	Normalised number of collaborations internal or external between institutions of each country.
F_OA	Leiden ranking	Normalised number of OA publications across the institutions of each country.
F_F_MF	Leiden ranking	Percentage of female researchers over the scientific population of institutions of a country.
F_industry	Leiden ranking	Normalised number of collaborations between the institutions of a country and industry.

4.4. Results

4.4.1. Combining RRI data with Leiden Ranking data

Figure 21 combines data from the Leiden Ranking[25] and MORRI Indicators dataset[26] and allows for visualisation of the correlations observed. The graph is coloured so each colour represents a country whilst each dot represents an institution. The clearest result here shows a strong correlation between engagement and policy (*r=0.79, n=344*). These data points both use SuperMoRRI data and show a strong association between the RRI policies adopted at a country level and the level of public engagement with science within that country.

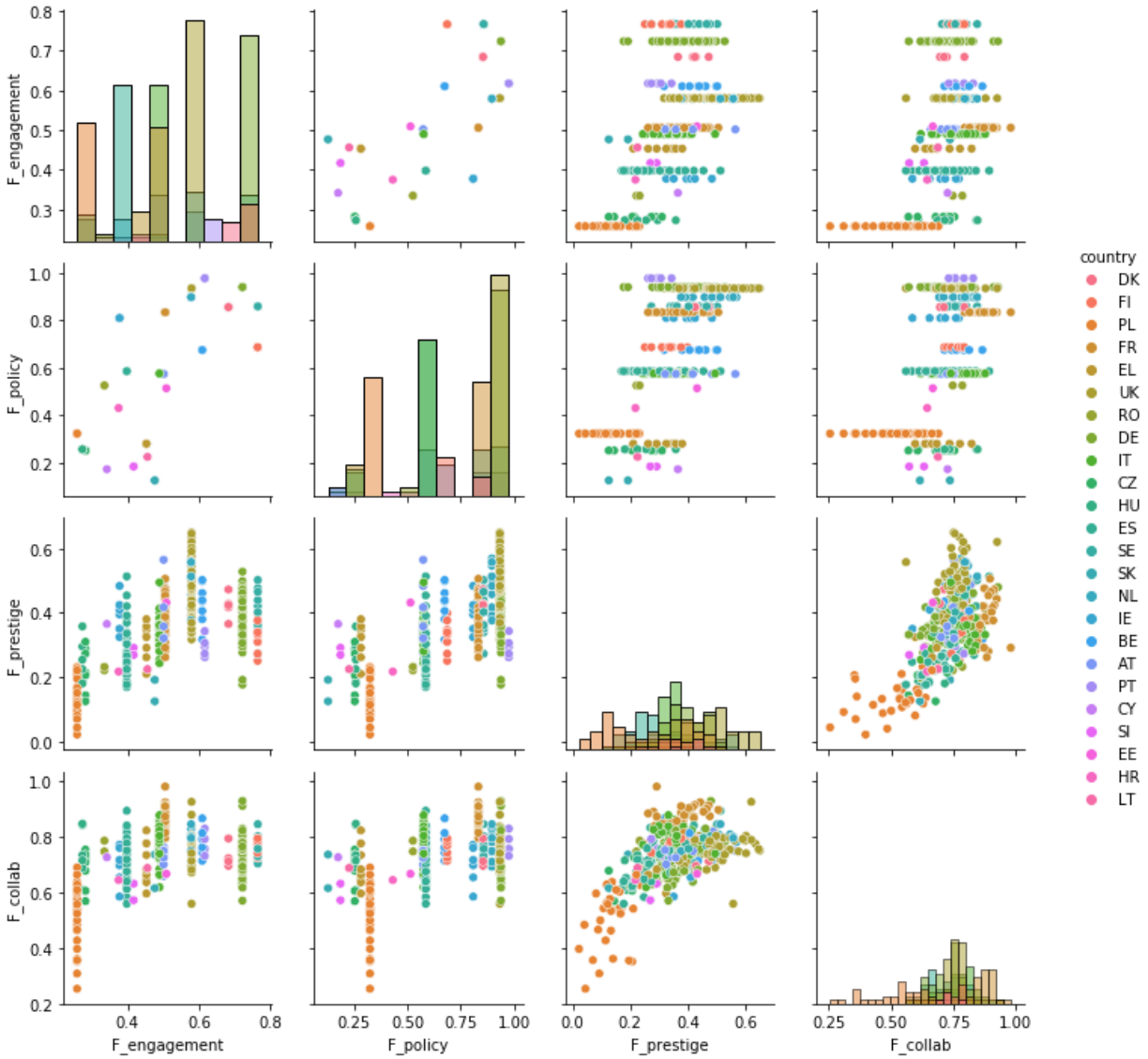

Figure 21: Pearson's r correlations between MoRRI and Leiden Ranking Data

[25]

[26] https://super-morri.eu/morri-2014-2018/

Using Pearson's r, Figure 22 shows a medium-strong correlation between institutional prestige and OA production (*r=0.63, n=344)* and also allows for the identification of country clusters by colour. This confirms our earlier results in Study 1 (section 3) which found a correlation between an institutions' rank and its levels of OA production (higher-ranking = more OA). It can be seen that UK institutions perform particularly well in terms of both institutional prestige and OA production. There is another conspicuous cluster in the lower portion of the graph indicating that Polish institutions do not in general perform well using these particular metrics.

There is a much closer alignment within the group for the remaining European countries. As discussed in the previous section, the UK situation is somewhat unique in Europe with the REF2021 Open Access mandate now fully implemented.

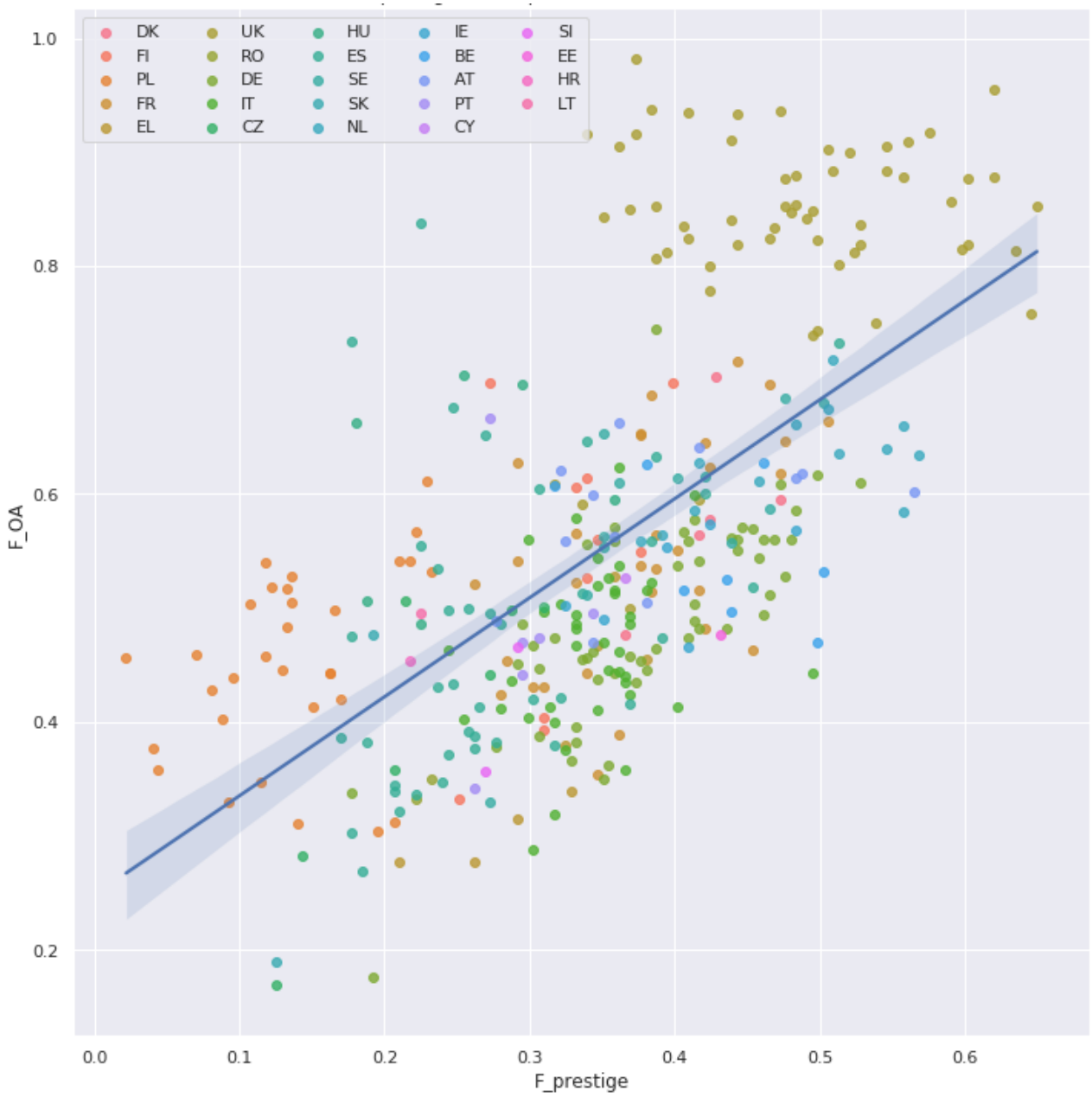

Figure 22: Institutional Prestige vs OA production

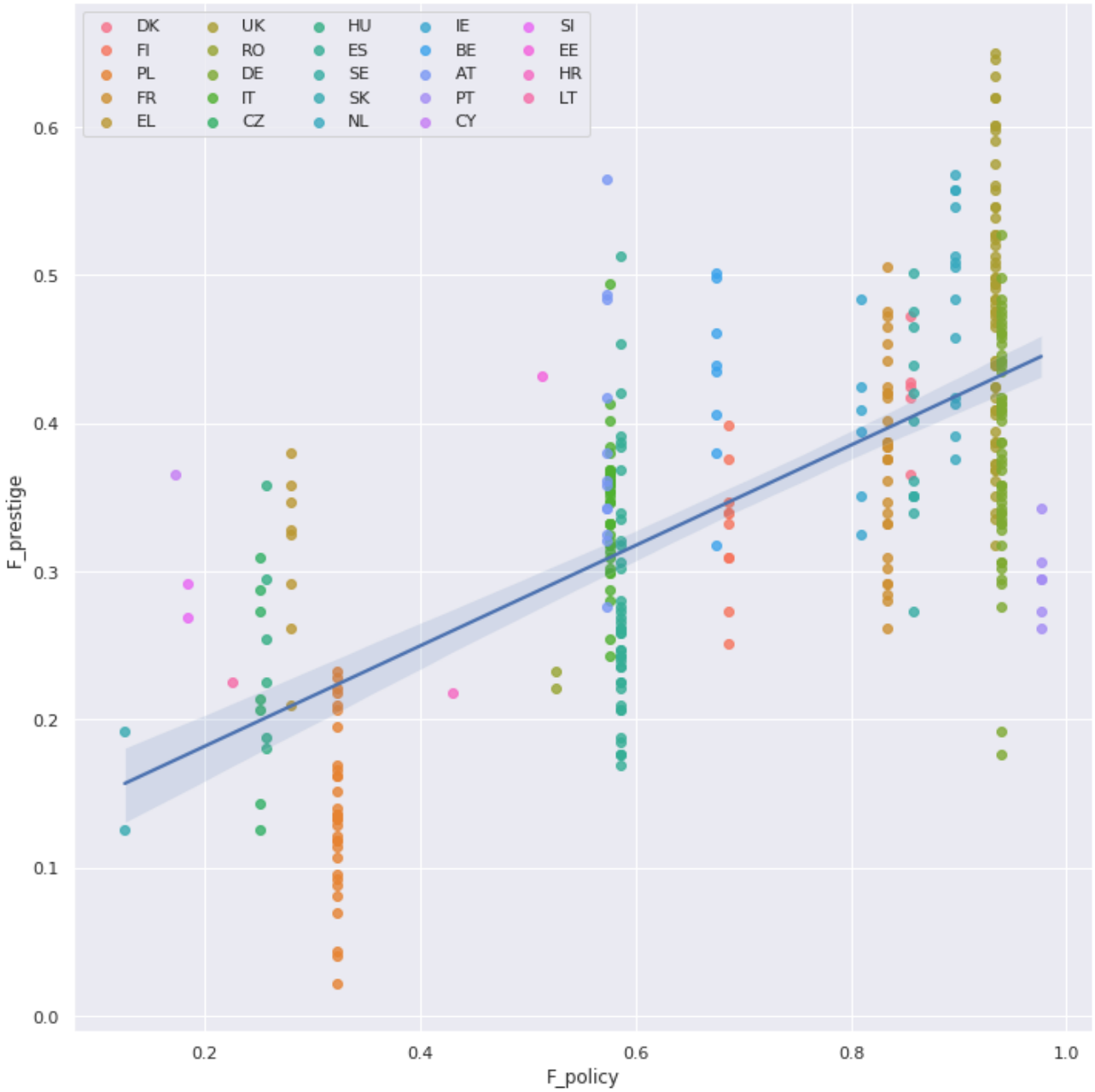

Figure 23: RRI Policy adoption vs Prestige

As can be seen in Figure 23, there is again a medium-strong correlation between the number of government mandated RRI policies and institutional prestige. The UK and Germany score highly in this regard, as do the Netherlands and Denmark. We may again relate this to ON-MERRIT's theme that the introduction of OS/RRI requires resources and so more well-resourced actors will be quicker to take it up.

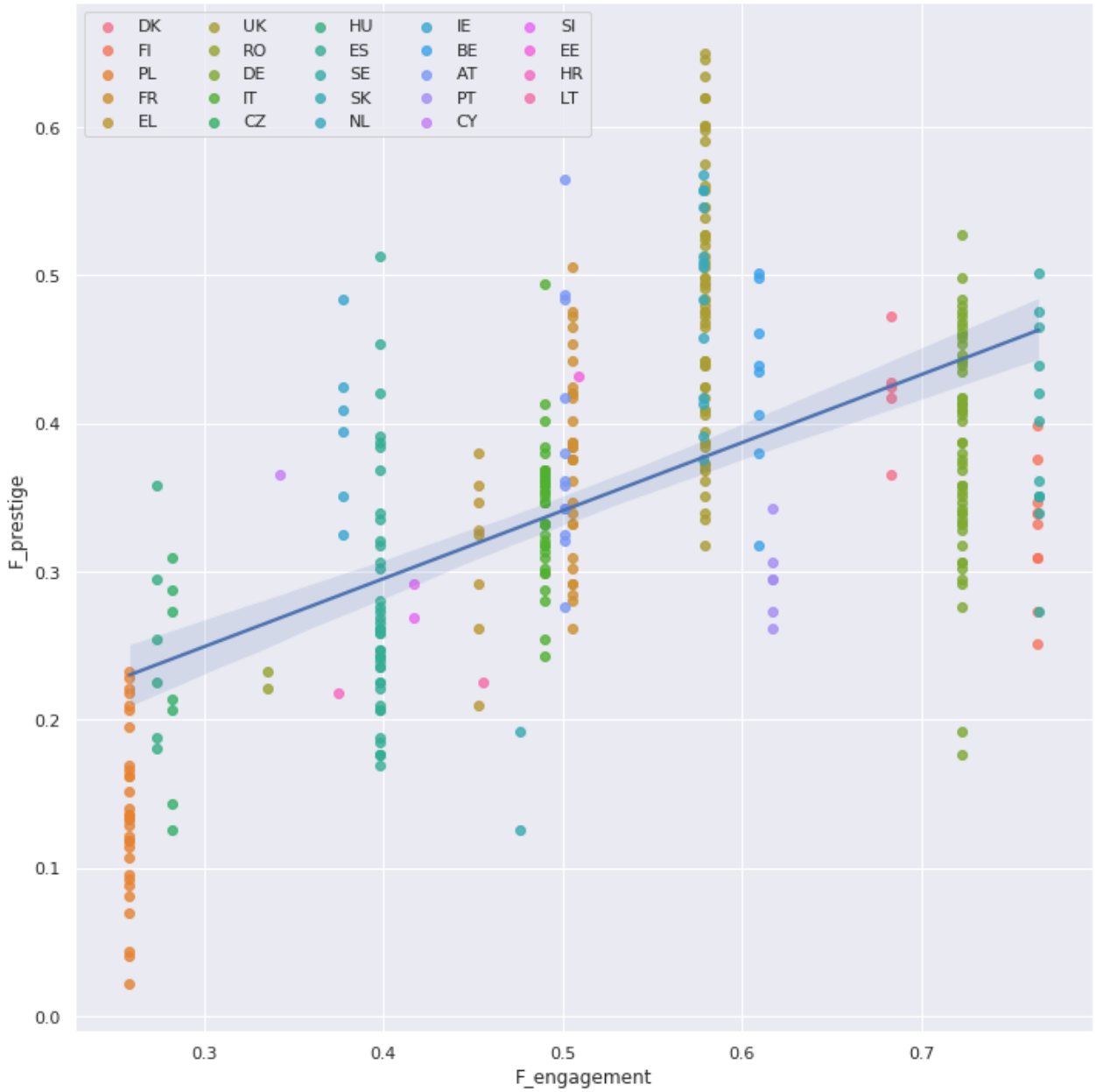

Figure 24: Public Engagement vs Prestige

There is a slightly lower correlation when one considers public engagement and prestige. This is one area where the UK does not perform as well as its European counterparts. It can be seen from Figure 24 that Germany, Denmark and the Netherlands are the leading European countries in regards to public engagement.

The RRI policy data from SuperMoRRI are focused on institutional and national policies and there is a medium-weak correlation between RRI policies and university collaborations. It does not consider the academic relationship with industry. It is not surprising therefore to find little correlation between RRI policies and industrial collaborations. One interesting finding is the lack of correlation between how a country performs in terms of gender equality policies (*F_gender*) and the actual balance in numbers of male / female researchers (*F_F_MF*). The data regarding gender is compiled from the SuperMoRRI data and does not specify figures for those identifying as "other" (e.g. non-binary).

Table 4: Correlations between Leiden Rankings (prestige), Collaboration (Leiden), Open Access Publishing (Leiden) and RRI Pillars shows all correlations for each of the data points in this study. The strongest correlation overall is for policy and public engagement. It is the correlations that can be observed when combining data from SuperMoRRI and the Leiden Ranking which have not been previously shown and offer new insight into how RRI policies and measures of prestige intersect.

Table 4: Correlations between Leiden Rankings (prestige), Collaboration (Leiden), Open Access Publishing (Leiden) and RRI Pillars

	F_gender	F_society	F_engagement	F_ethics	F_policy	F_prestige	F_collab	F_OA	F_F_MF	F_industry
F_gender	1.000	0.524	0.084	-0.011	0.224	0.116	-0.166	0.443	**0.084**	-0.006
F_society	0.524	1.000	0.122	0.147	0.314	0.061	-0.077	0.384	0.045	-0.167
F_engagement	0.084	0.122	1.000	0.190	**0.794**	0.576	0.398	0.248	-0.448	0.396
F_ethics	-0.011	0.147	0.190	1.000	0.300	0.318	0.044	0.404	-0.075	0.039
F_policy	0.224	0.314	**0.794**	0.300	1.000	**0.675**	0.454	0.480	-0.374	0.285
F_prestige	0.116	0.061	0.576	0.318	**0.675**	1.000	**0.610**	**0.632**	-0.236	0.399
F_collab	-0.166	-0.077	0.398	0.044	0.454	**0.610**	1.000	0.348	-0.071	0.324
F_OA	0.443	0.384	0.248	0.404	0.480	**0.632**	0.348	1.000	-0.073	0.210
F_F_MF	**0.084**	0.045	-0.448	-0.075	-0.374	-0.236	-0.071	-0.073	1.000	-0.393
F_industry	-0.006	-0.167	0.396	0.039	0.285	0.399	0.324	0.210	-0.393	1.000

4.4.2. RRI - Country comparisons using Igloo plots

The diagrams in the following section were generated using the Web-Igloo tool (Kuntal, Ghosh, and Mande 2014) for displaying multivariate data. Web-Igloo visualizes multivariate data in a 2D chart of multiple quantitative variables represented as anchors on a semicircle. This tool identifies clusters in the data and also enables the ability to see the features responsible for the clustering. The projected data points are mapped to class labels using a simple metadata file.

The strength for each data point is indicated by the inverse graph at the top of each figure. The total area for all data points is bounded and coloured to allow us to see an individual country's total weighted performance.

Figure 25: Igloo Plot for United Kingdom further demonstrates the UK's prominent position as a producer of OA research literature. The UK also scores highly for ethics and the impact of research on society but fairs less well for collaborations with industry. It was noted by Frenken, Heimeriks and Hoekman (2017) that "*UK universities do particularly well in citation impact and internationalization, but are rather poor at industry involvement."* The results here further strengthen that observation. This data also reveals that, whilst the UK has numerous policies to promote gender equality, male researchers still outnumber their female counterparts overall.

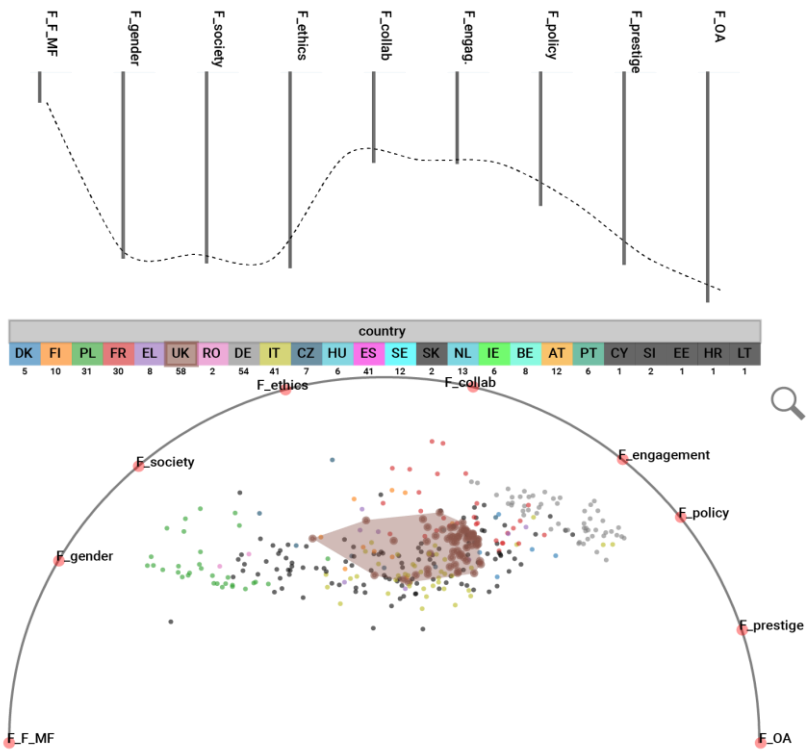

Figure 25: Igloo Plot for United Kingdom

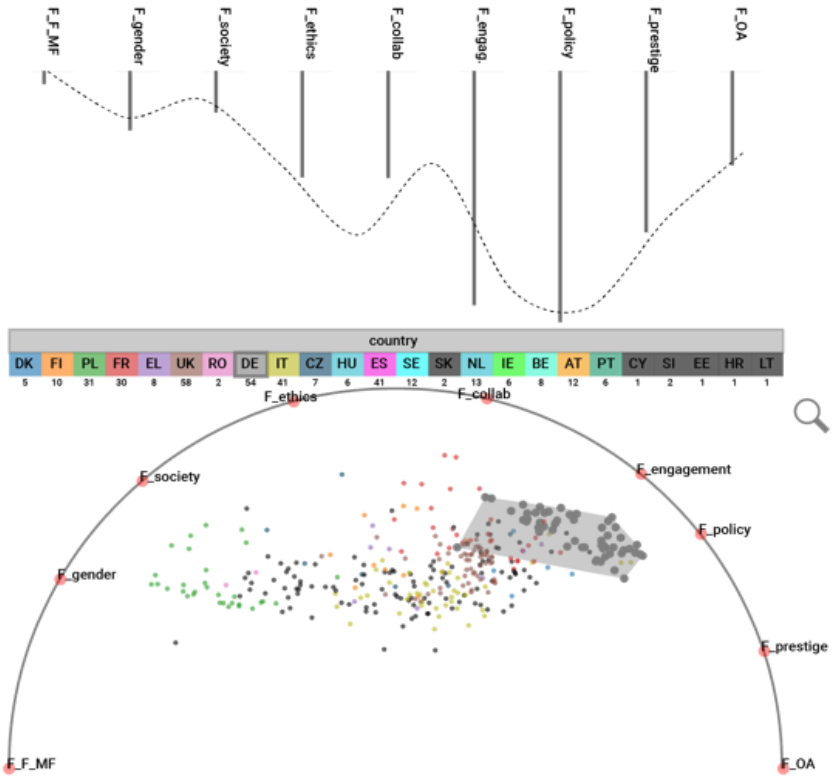

Figure 26: Igloo Plot for Germany

The data for Germany, Figure 26: Igloo Plot for Germany, paints a very different picture of the RRI and OA landscape. Overall, there are more RRI policies applied than in the UK; however, this is not reflected in the amount of OA research literature produced by German institutions. Further, Germany scores much higher than the UK for public engagement with science and about equally in terms of collaborations with researchers from other institutions. Overall however, German institutions do not score as highly as UK institutions in terms of international prestige, based on data from the Leiden Ranking. The under-representation of female researchers can also again be seen.

We lastly look at the state of play in two other European countries as both are outliers in one particular area in the application of RRI policies. Using data from the Leiden Ranking, it can be seen that the Czech Republic Figure 27: Igloo Plot for Czech Republic has a strong balance of male / female researchers (*F_F_MF)*, closer than most other European countries. However, the country scores extremely poorly in terms of the adoption of gender-focused RRI policies and the country is ranked 26 out of 28 in terms of the gender pay gap. (Figure 27)

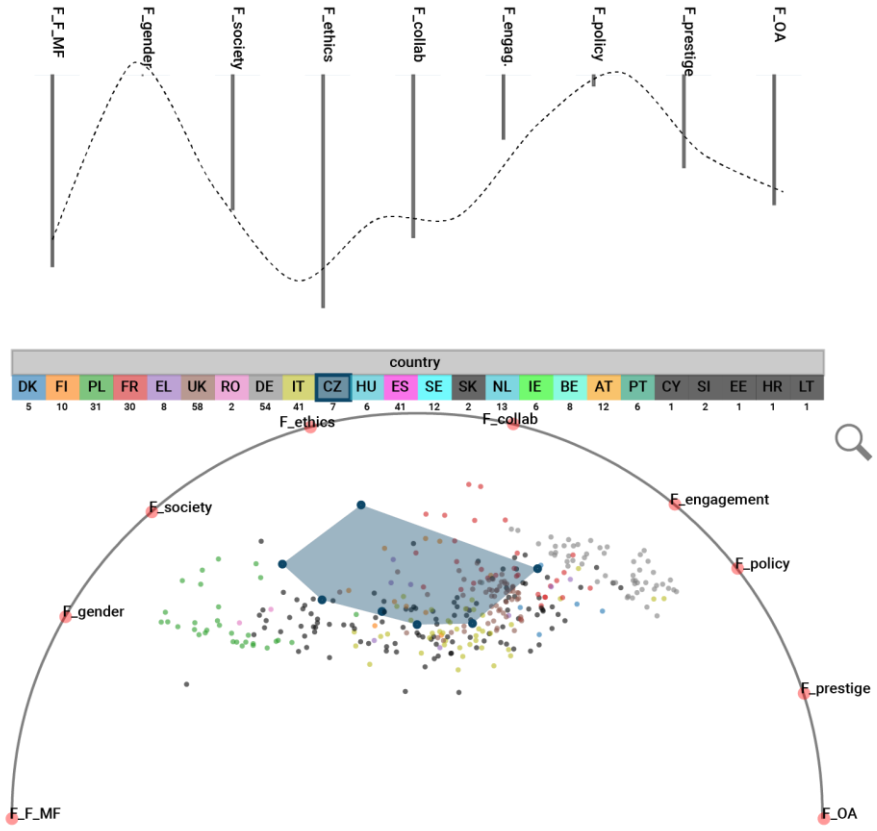

Figure 27: Igloo Plot for Czech Republic

The lack of RRI policies in place in the Czech Republic demonstrates that the fairly equitable gender balance has not been achieved through the adoption of policy but is merely a reflection of the current research landscape.

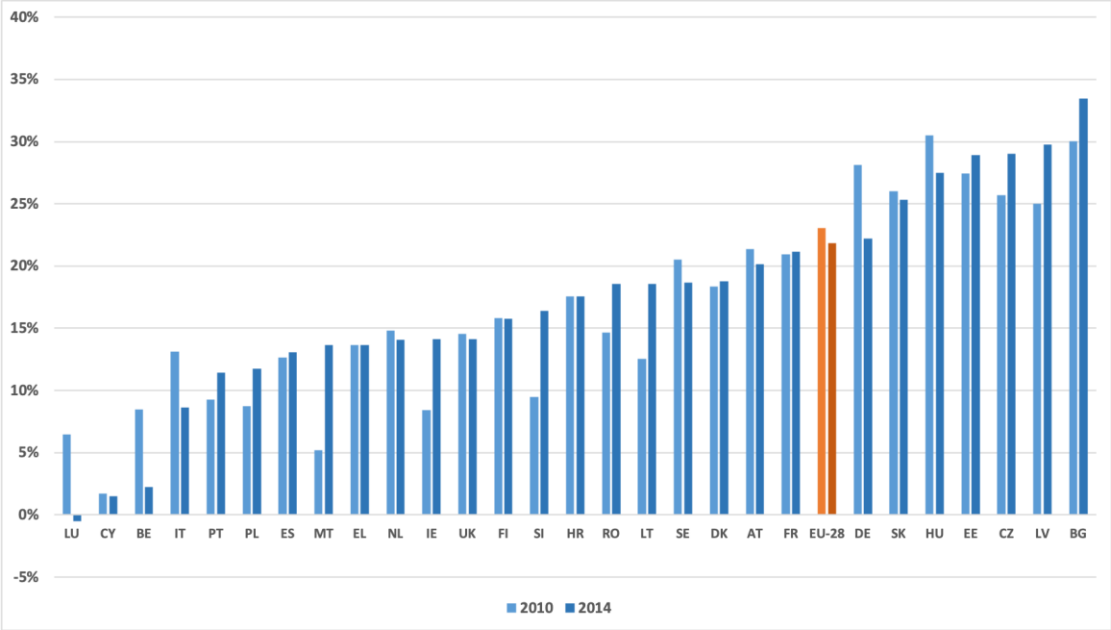

Figure 28: Gender pay gap in European countries. Source: SuperMoRRI data

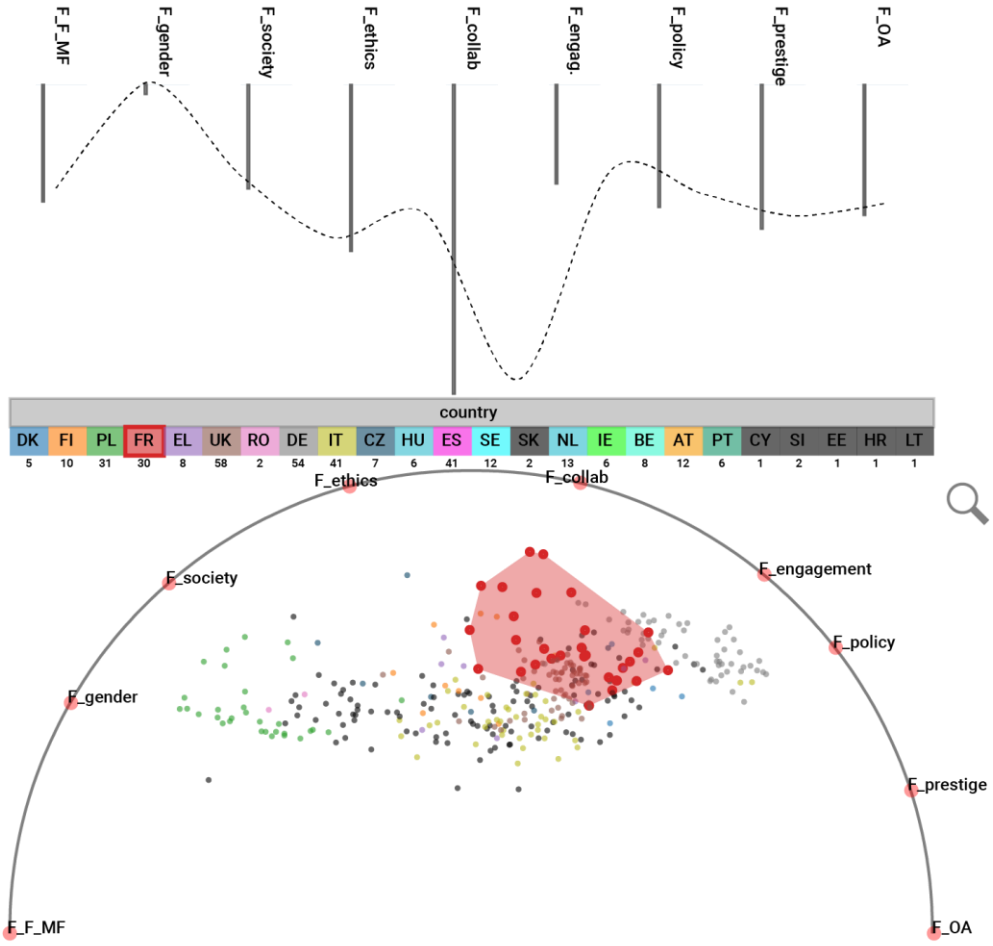

Figure 29: Igloo Plot for France

Finally, in Figure 29, we examine the case for France, which scores lower than most of its European counterparts for many of the features examined, scoring particularly poorly for gender equality policies. Further, despite a correlation between collaboration and prestige being identified in the previous section, France as a whole does not appear to benefit overall from its undoubtedly high level of collaboration.

4.5. Discussion

There has undoubtedly been much progress in the development of new RRI indicators since their inception as a part of the EC-Funded MoRRI project. The data encompassed by the MoRRI indicators provide a wealth of evidence for the adoption and impact of RRI policies across Europe. Our Igloo plots for this data show clear and very significant differences across European countries when viewed at the country level. The new EU (EU13) countries in particular score poorly for many of the 36 indicators with especially noticeable issues in regards to gender equality and diversity.

Our results here demonstrate a strong overall correlation between one of the RRI pillars, measures of *public engagement* with science, and RRI policies at the national level (*r=0.79, n=344)*. We see demonstrably higher levels of public engagement with science in countries where these policies are more embedded. Further, we show a medium to strong correlation (*r=0.67, n=344*) between institutional prestige and OA production which aligns with our earlier findings, that the OA landscape is clearly divided between prestigious and non-prestigious institutions. It can be seen that UK institutions perform particularly well in terms of both institutional prestige and OA production. But in terms of the latter, its relative OA production rate is lower than many other European countries as shown in the Study 2.

A further interesting finding is the lack of correlation between how a country performs in terms of gender equality policies (*F_gender*) and the actual balance in numbers of male / female researchers. (*F_F_MF*). This is particularly noticeable in Portugal, Bulgaria, Latvia, Lithuania and Hungary where the balance of male / female researchers is approximately equal. As a comparison, for France the split in 2018 was 35.9% female, and for Germany, 36.2% female. The situation in each individual country is therefore more convoluted and complex with social drivers and the status quo appearing to be powerful influencers.

Combining Super MoRRI indicators with ranking data from Leiden allowed us to measure for the first time the link between RRI factors and institutional prestige, finding a medium-strong correlation for this. Each of our results continue to highlight that it is the higher ranked, more prosperous and more prestigious institutions that appear best able to adopt, adapt to, and benefit from, the evolving Open Science and RRI landscape.

However, our results also indicate that policies are not always linked to practices in linear ways, and that individual indicators taken in isolation could be misleading. The Czech Republic, for example, has a fairly equitable gender balance in terms of numbers of male/female researchers. Yet this is not due to specific policies, and in addition this equity in terms of numbers is not mirrored by equity in terms of the gender pay-gap.

5. The structure of knowledge production in research on three key UN Sustainable Development Goals

5.1. Introduction

Following the above discussions, the following two studies will focus in on research that is being conducted on three key UN Sustainable Development Goals (SDGs). This line of inquiry is motivated by ON-MERRIT's aim of providing new evidence on research contributing to these crucial goals. In 2015, the United Nations adopted the SDGs, a collection of 17 global goals set by the United Nations General Assembly to provide guidance and targets for global, stakeholder-led sustainable development by 2030.[27] The 17 goals cover a wide array of often interconnected issues, as can be seen in Figure 30 below. These 17 goals relate to 169 individual targets, and progress towards these targets is to be monitored by over 230 indicators (this number continues to grow, and many more have been proposed) (Zinkernagel, Evans, and Neij 2018).

Figure 30: UN Sustainable development goals

In accordance with ON-MERRIT's mission, we focus on three key SDGs:

- SDG 2 Zero Hunger: End hunger, achieve food security and improved nutrition and promote sustainable agriculture
- SDG 3 Good Health and Well-Being: Ensure healthy lives and promote well-being for all at all ages
- SDG 13 Climate Action: Take urgent action to combat climate change and its impacts

[27] https://sdgs.un.org/

In this study, we establish baselines in terms of general publication and collaboration patterns on which the following study on stratification in OA publishing will build. We focus on key factors that structure the academic system of knowledge production (Zuckerman 1988; J. R. Cole and Cole 1973): researchers' ages and genders, institutional standing, and international collaboration.

5.2. Background

The SDGs cover a range of pressing issues. Eradicating poverty and hunger, ensuring access to clean water, quality education, establishing good health and well-being, as well as gender equality, and combating the intensifying climate crisis, amongst others, are all urgent issues that the international community wants to address. Given the scientized nature of today's societies (Drori et al. 2003), research and innovation is expected to make a substantial contribution in meeting these challenges (see also Bautista-Puig et al. 2021). It is therefore timely to explore how research on crucial areas for sustainable development has evolved over the last two decades.

Our key concerns are how key trends of equity in science relate to research conducted on the three target SDGs. How is institutional prestige related to the production of research on the three SDGs? How is individual publication output related to academic age? Do men publish more than women, and by how much? To anchor the following analyses, we review available evidence on general sociological trends in scholarly communication below.

5.2.1. Stratification in scholarly communication

There is strong evidence that women are under-represented in academic publishing. Lariviere et al. (2013, 212) find that among articles published between 2008 and 2012 and indexed in the Web of Science databases, "*women account for fewer than 30% of fractionalized authorships*". This is in line with West et al. (2013), who similarly find the share of female authorships to be slightly below 30% for data from JSTOR in the period 2000-2009. Although female participation in terms of authorship has risen significantly, from about 10% for the period 1900-1960 (West et al. 2013, 2) to the aforementioned 30%, there remains work to do to achieve equity. Moreover, inequality still persists in other aspects: women are still underrepresented in terms of first authorships (Larivière et al. 2013; West et al. 2013) and last authorships in fields where the last author position signifies prestige and seniority (West et al. 2013). Furthermore, there are substantial gender differences between disciplines. For the period 1990-2011, West et al. (2013, 2) report a share of female authorships below 15% for the fields of Mathematics, Philosophy and Economy, but a share above 40% for the fields of Education, Demography and Sociology. Similarly, Larivière et al. (2013, 212) found higher shares of female authorships in fields associated with "care", such as nursing, education, and social work, and lower shares in "*high-energy physics, mathematics, computer science, philosophy and economics*". It is therefore reasonable to expect a higher share of female authorships for SDG Health/Well-Being (SDG 3), but a lower share of female authorships for SDG Climate Action (SDG 13), given the former's focus on care-related issues, and the latter's association with physics and computer science.

Another determinant of research productivity is academic age. A comprehensive body of literature (see e.g., S. Cole 1979; Fox 1983; Costas, Leeuwen, and Bordons 2010; Yair and Goldstein 2020; Rørstad and Aksnes 2015) suggests that research productivity follows an inverted U-shaped distribution, with high productivity in the middle of researcher's careers, and low productivity at its start and end. These general patterns differ

among disciplines. Rørstad and Aksnes (2015) found no decrease in productivity, but a general upward trend in line with age groups for the social sciences and humanities. Similarly, productivity increased in medicine, with a slow initial increase and a later plateau. Researchers from the natural sciences showed a slight drop in productivity in their later years, with the largest drop found for engineering and technology. However, reviewing evidence on whether scholars' highest impact papers are published earlier or later in their careers, Fortunato et al. (2018) found no effect. Therefore, while *research productivity* seems linked to academic age, *research impact* seems not to be.

Affiliation has been found to be another important factor associated with research productivity. Investigating whether prestigious university departments enable research productivity or whether they just hire the most productive researchers, Allison and Long (1990) found clear evidence for the former: researchers' productivity and impact rises when scholars move to more prestigious institutions, and falls if they are downwardly mobile. Allison and Long suggest multiple plausible mechanisms related to cumulative advantage (Ross-Hellauer et al. 2021) to explain this effect, from better facilities to greater visibility of research, to higher motivation and intellectual stimulation at more prestigious institutions. In addition, differences in the emphasis placed on teaching and research, as well as the overall extent of administrative duties could also contribute to the observed effect.

5.2.2. Mapping research to UN Sustainable Development Goals (UN SDGs)

Research interest on mapping research to the UN SDGs has grown rapidly over recent years. A common approach is to use keyword searches in academic databases for terms like "sustainable development goals" (see e.g., Gonzalez Garcia, Colomo Magana, and Civico Ariza 2020; Meschede 2020; Nazari et al. 2020; Pizzi et al. 2020; Sweileh 2020). This approach, however, is only able to find research that explicitly mentions the UN SDGs, which omits relevant research which does not clearly identify itself as such and includes meta-research about the SDGs themselves which is out of scope. To find research that is applicable to the SDGs, while not directly mentioning them, approaches rely on extensive lists of keywords that map to terms related to specific SDGs. Scopus developed such lists for all SDGs (Jayabalasingham et al. 2019) and now provides these predefined queries via their interface[28]. A conceptually similar approach is that of OSDG[29], which uses an extensive set of expert-developed ontologies, which were merged into an integrated ontology. The terms from the integrated ontology were then mapped to Fields of Study (FOS) from MAG via the Levenshtein distance (Pukelis et al. 2020).

A known challenge in finding research that relates to the mapping of UN SDGs to research output is the operationalisation of the concrete terms (Armitage, Lorenz, and Mikki 2020). Is a publication sufficiently relevant to SDG Climate Action (SDG 13) if it mentions "climate change", or does it need to mention climate change in relation to action tasks such as "mitigation" or "adaptation"? Is the inclusion of the term "marine" sufficient for a publication to qualify for SDG 14 (life below water) or not? Armitage, Lorenz, and Mikki (2020) found the approach by Scopus to be very wide in scope, and it can be expected that the approach by OSDG is similar in terms of scope. All these challenges notwithstanding, finding and analysing research that is

[28] https://blog.scopus.com/posts/sustainable-development-goals-sdgs-on-scopus
[29] https://osdg.ai/

relevant for tackling the UN SDGs is of high relevance, given that scientific research is expected to combat these key issues (Bautista-Puig et al. 2021).

5.3. Methods

5.3.1. Mapping publications to UN SDGs

To map publications from MAG to the three target SDGs Zero Hunger (SDG 2), Health/Well-Being (SDG 3) and Climate Action (SDG 13), we used the mappings between MAG Fields of Study (FOS) and SDGs by the OSDG.ai project (see section 5.2.2).

Since this approach is prone to including unrelated papers, we took several mitigating steps. First, we considered the hierarchy of MAG FOS. FOS from MAG are organised hierarchically, with the top level comprising general fields such as biology, medicine, philosophy, history, physics, or business[30]. To avoid the inclusion of unrelated papers, we only kept papers where the FOS from the mapping was on level 3 (out of 6) or lower, thus excluding papers only mapped to broader top-level fields. Second, MAG includes a "score" for the confidence of the mapping between a given paper and a FOS. Although the documentation states that values can range from 0 to 1, 80% of mappings between papers and FOS have a score between 0.32 and 0.56, with the top 25% of mappings from papers to FOS above a score of 0.4972. We only included papers where the FOS from the OSDG ontology had a confidence score of 0.497 or higher, thus retaining mappings which fall in the top 25% of empirical confidence scores within the database. Finally, we restricted the dataset to include publications from 2006 until 2020. The upper bound was determined by the limits of the available data, while the lower bound was motivated by keeping a similar time window as in Study 4 (section 6), where we investigate the development of OA publishing.

To assess the validity of the approach, we randomly sampled 100 papers, 33 for SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13) each, and 34 for SDG Health/Well-Being (SDG 3). This exercise intended to establish whether the approach was credible, or led to the frequent inclusion of unrelated publications. For each paper, we assessed the title, abstract, and in some cases the full paper, to establish whether the paper could reasonably be understood as referring to the respective SDG. A high share of sampled publications from SDG Health/Well-Being (SDG 3) were assessed as relevant to the SDG (91%, 90% CI: 78%-97%), with a moderate share (79%; 90% CI: 63%-89%) of sampled publications from SDG Climate Action (SDG 13) seeming actually relevant, but a lower share (61%, 90% CI: 45%-75%) for SDG Zero Hunger (SDG 2). In most cases, papers that were not relevant to SDG Zero Hunger (SDG 2) seemed more relevant to the closely-related SDG Health/Well-Being (SDG 3). This interrelatedness of the SDGs is also reflected in the overlap between papers sampled based on the three SDGs. Figure 31 depicts the number of papers sampled from each SDG, given a sub-sample of the total sample. Among 1 million papers in the sub-sample, the majority exclusively relates to SDG Health/Well-Being (SDG 3) (89.69%), 6.8% relate to SDG Zero Hunger (SDG 2), and 3.06% relate to SDG Climate Action (SDG 13). In terms of overlaps, the highest overlap is between SDGs Zero Hunger (SDG 2) and Health/Well-Being (SDG 3), where 1553 publications were both sampled as referring to SDGs Zero Hunger (SDG 2) and Health/Well-Being (SDG 3) respectively (0.15% of all papers in the subsample). Overlaps between the other SDGs were substantially lower.

[30] https://academic.microsoft.com/topics

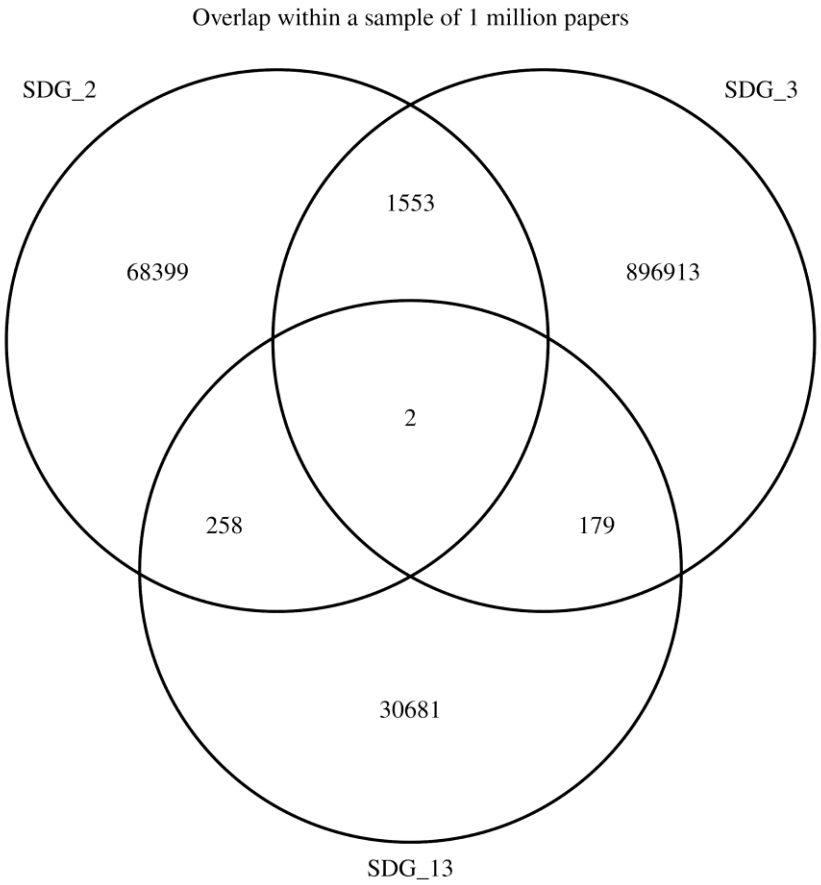

Figure 31: Overlap between publications sampled from the three UN SDGs

5.3.2. Assessing author gender

There are various approaches to obtaining gender information for authors of academic publications, ranging from semi-automatic to fully automatic approaches (King et al. 2017; Larivière et al. 2013; Thomas et al. 2019). Given the size of our data, we pursued an approach based on the genderize.io API, which has been used in many previous studies (e.g., Dion, Sumner, and Mitchell 2018; Iefremova, Wais, and Kozak 2018; Hart, Frangou, and Perlis 2019; Thomas et al. 2019; Olejniczak and Wilson 2020). We here detail the exact steps and parameters chosen.

Author data from MAG includes a column for normalized author names (with punctuation removed, normalised first and last names based on cultural norms, and converted to lowercase). Our dataset included 13,999,384 uniquely identifiable authors. From the column of normalized names, we extracted the first word component with at least two characters[31]. Through this approach, we were able to extract 11,474,798 *potential first names* (81.97% of all author names). Many of the potential first names were duplicates, resulting in 509,731 unique first names. Using the genderizeR package (Wais 2006; Wais et al. 2019), we queried all unique names against the genderize.io API. Mapping the genderized unique names to all authors, genderize.io returned a result for 96.38% of *potential first names*.

The results from genderize.io provide data on the probability of a gender mapping, as well as a count of how frequent a given name is in the genderize.io database. We used both data points to improve the reliability of the gender assignment (Wais 2016). In other studies using genderize.io, there is no clear consensus on which

[31] We retained the first capture group from the regular expression "(\w{2,}?)\s"

thresholds to use however (Dion, Sumner, and Mitchell 2018; Iefremova, Wais, and Kozak 2018; Hart, Frangou, and Perlis 2019; Thomas et al. 2019). Since our core goal was to have precise estimates for the gender split in OA publications, we opted for a higher precision, with potentially lower recall. We chose to only include names that appeared at least five times in the genderize.io database, and kept assigned genders only where the API returned a probability for a correct assignment of at least 85%. We therefore were able to obtain data on author genders for 59.78% of all authors, 72.93% of authors with a potential first name, and 75.67% of all authors for which we had a response from genderize.io. These fractions are very similar to other studies with large datasets (King et al. 2017; Larivière et al. 2013).

5.3.3. Field normalization

For comparing metrics such as the number of citations across fields, normalization is necessary. MAG provides fields of study which are generated algorithmically in considering semantic similarity (Wang et al. 2020; 2019). Previous research has found the hierarchy of MAG FOSs to be too incoherent to be suitable for normalization (Hug, Ochsner, and Brändle 2017; Waltman and van Eck 2019). We therefore follow Hug et al. (2017) and normalise citations by venue and year. For each journal we calculate the average number of citations towards the journals' papers per year. For conferences, MAG uniquely identifies single "instances" of a conference, for which we calculate the average citation count as well. The actual citation count of any given paper is then divided by the corresponding number of its journal or conference. Thus, if a paper published in 2015 has received 10 citations to date, while the average of citations to papers published in 2015 in the same journal is 5, the paper gets a standardized citation value of "2". This means that the paper was cited twice as much as the average paper in this journal and year. A known limitation of journal-based normalization is that it might advantage authors publishing in journals which tend to accrue fewer citations on average, whereas the basis for denomination is much broader when field-based normalization is used[32]. For this reason, results based on normalised citations should be interpreted with some caution.

5.4. Results

5.4.1. Who publishes research on the SDGs zero hunger, good health and well-being and climate action?

General overview

The number of publications that can be mapped to the three SDGs Zero Hunger (SDG 2), Health/Well-Being (SDG 3) and Climate Action (SDG 13) is increasing between 2006 and 2019 (Figure 32) Using the mapping from OSDG to MAG FOS, we find the sample on SDG Climate Action (SDG 13) encompassing 11,258 to 27,682 yearly publications, SDG Zero Hunger (SDG 2) encompassing 24,977 to 55,343 yearly publications, and SDG Health/Well-Being (SDG 3) encompassing 380,225 to 718.,00 yearly publications. These numbers correspond to growth factors over the whole period of 2.4 for SDG Climate Action, 2.2 for SDG Zero Hunger, and 1.9 for SDG Health/Well-Being, which equals average yearly growth rates of 5-7%.

[32] We thank Thed van Leeuwen for raising this concern.

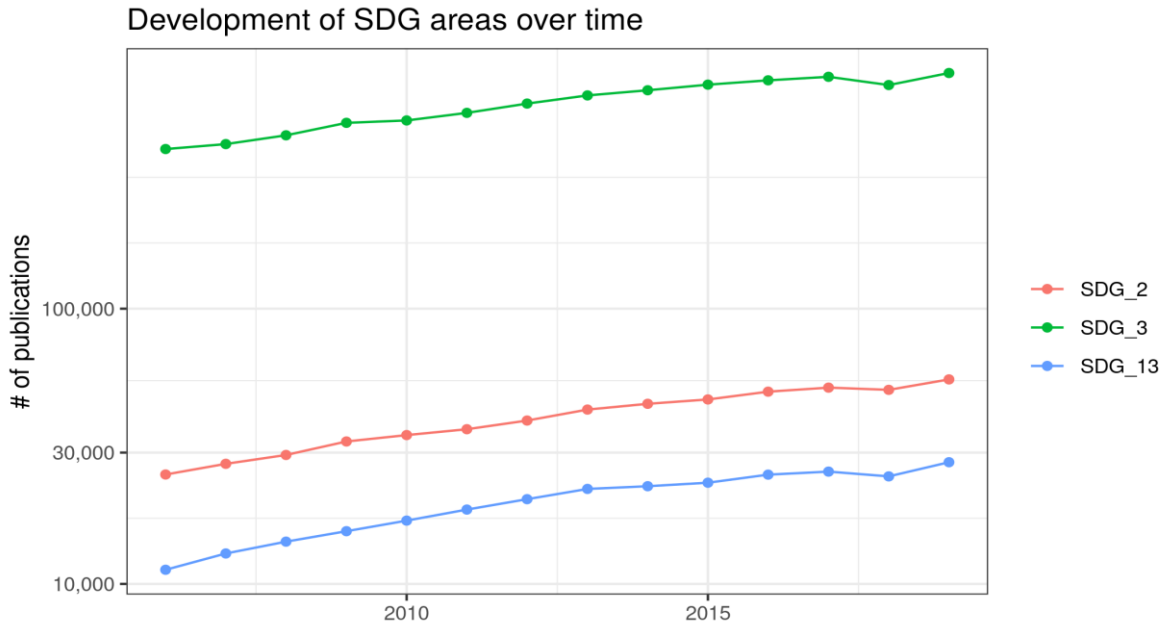

Figure 32: Numbers of yearly publications in the SDG areas over time

While the number of research articles published in these three areas has been rising, the average and standardised number of total citations received by more recent publications is lower than for older publications (Figure 33). This effect cannot be explained by the time-lag inherent in citations, since citations are normalised dividing the number of citations a given publication has received by the average number of citations of all publications from the same venue (journal or conference) and year[33]. The observed effect might therefore represent a genuine trend in itself.

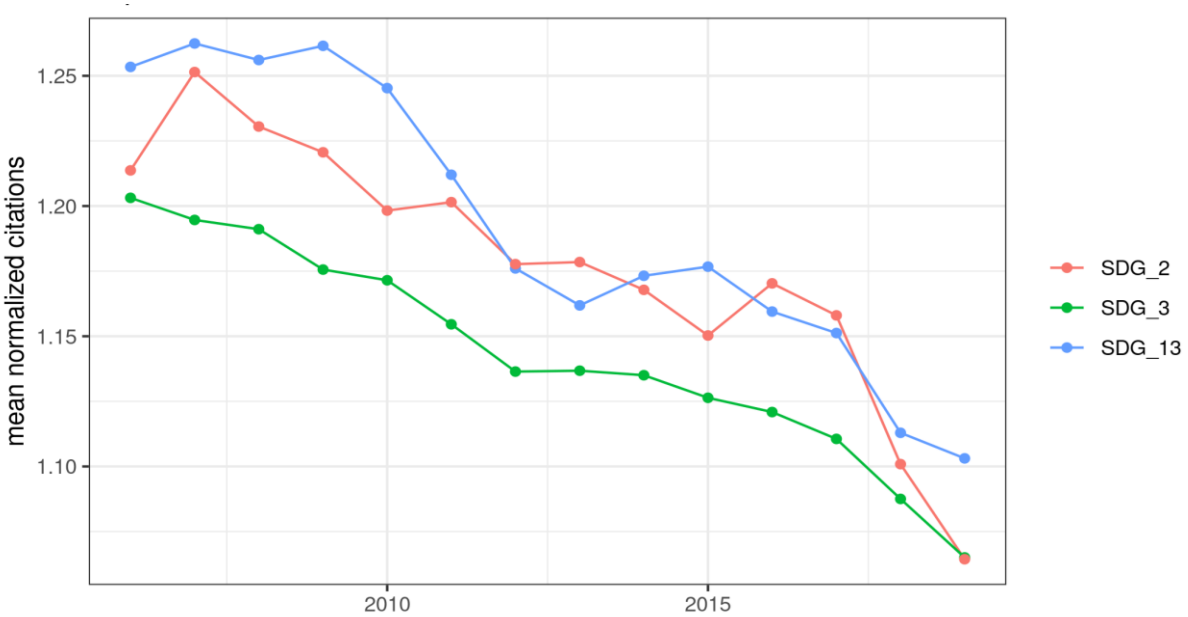

Figure 33: Impact of SDG research over time

[33] See also the following notebook demonstrating the standardisation procedure: https://github.com/on-merrit/sdg_analysis/blob/main/notebooks/06-standardisation.pdf

Institutional prestige

Investigating the role of institutional prestige, we find strong correlations with the number of publications and citations produced and received by the various institutions. Institutions that are included in the Leiden Ranking have higher numbers of citations and publications across all three SDGs (Figure 34) This result is to be expected, given that a substantial number of "core publications" is a precondition for an institution's inclusion in the Leiden Ranking (see section 2.2).

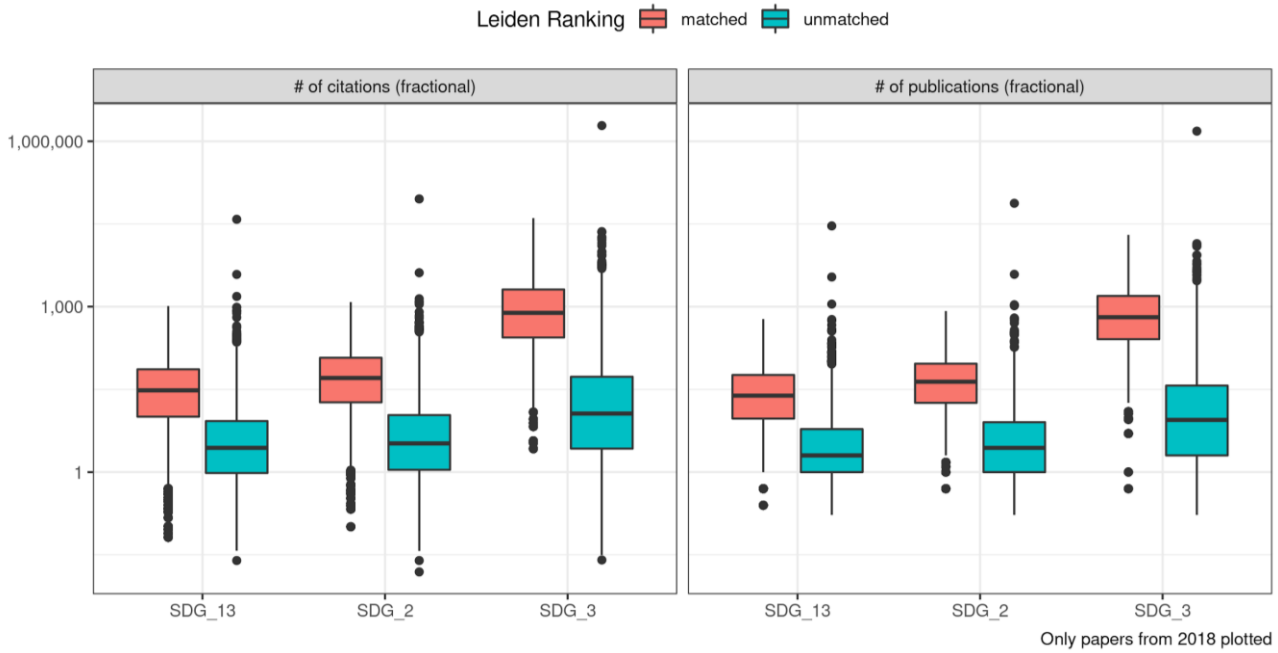

Figure 34: Numbers of citations and publications of institutions split by inclusion into the Leiden Ranking. The figure depicts the numbers of citations from and publications to institutions which are or are not ranked in the Leiden Ranking, split by to which SDG

When considering only institutions within the Leiden Ranking, we find a strong association between an institution's $P_{top\ 10\%}$[34] and the numbers of publications produced and citations received within the three SDG areas (Figure 35) The level of an institution's overall research production and impact therefore is closely linked to its output and impact within the three SDG areas. Comparing the number of papers published by authors from a given institution (fractional counting) in 2018 with the institution's ranking position (2015-2018), we find moderate to strong positive correlations. The association is strongest for SDG Health/Well-Being (SDG 3) (r = .81), followed by SDG Climate Action (SDG 13) (r = .63) and SDG Zero Hunger (SDG 2) (r = .47). This stratification is very stable over the period 2008-2018 (Figure 36).

[34] See Chapter 2.2. for an explanation of the indicator.

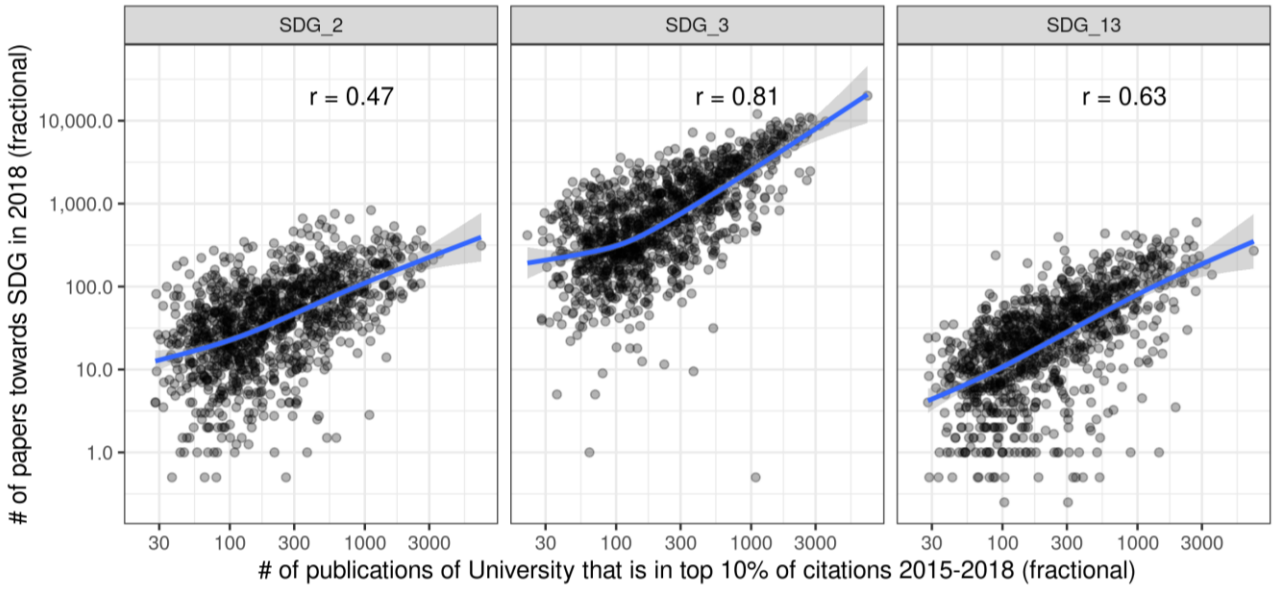

Figure 35: Correlation between $P_{top\ 10\%}$ (Leiden Ranking) and total paper output per SDG

Considering SDG Health/Well-Being (SDG 3) first, the top quintile (top 20% percent) of institutions according to the Leiden Ranking account for more than 50% of the publications, while the bottom two quintiles account for just 5-10% of publications each. This overall pattern is equally present in SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13), however its extent differs. While SDG Climate Action is fairly similar to SDG Health/Well-Being, the gap between low-ranking and high-ranking institutions is much smaller in SDG Zero Hunger. About 40% of publications can be attributed to the top 20% of institutions in SDG Zero Hunger (SDG 2), with all other percentile groups having bigger shares of the overall production of publications compared to SDG Health/Well-Being (SDG 3) and SDG Climate Action (SDG 13).

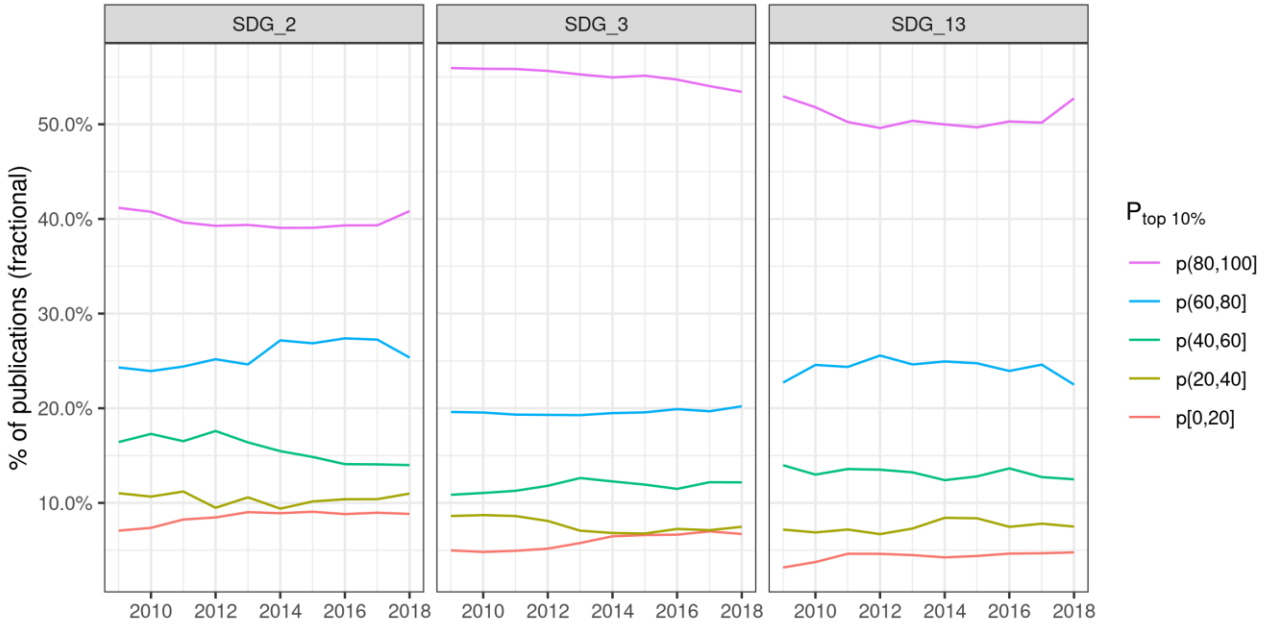

Figure 36: Share of fractional publications by quintiles of the distribution of $P_{top\ 10\%}$

Considering the shares of citations per institution, we find similar, if not stronger, stratification (Figure A1). Publications from the top 20% of institutions accrue 59-61% of citations in SDG Health/Well-Being (SDG 3),

about 52-55% of citations in SDG Climate Action (SDG 13) and slightly above 42-44% of citations in SDG Zero Hunger (SDG 2). Again, these differences are very stable over the considered time window.

Academic age

Moving from the institutional to the individual level, we investigated academic age as a proxy for the seniority of individual authors. In line with overall trends in scientific publishing (see e.g. Costas and Bordons 2011), we find clear differences in author ages between author positions on the byline (Figure 37). Across all three SDGs, first authors have the lowest academic age (mean age of 9-12 years), last authors the highest age (mean age of 16-19 years), with middle authors in between (mean age of 12-14 years).

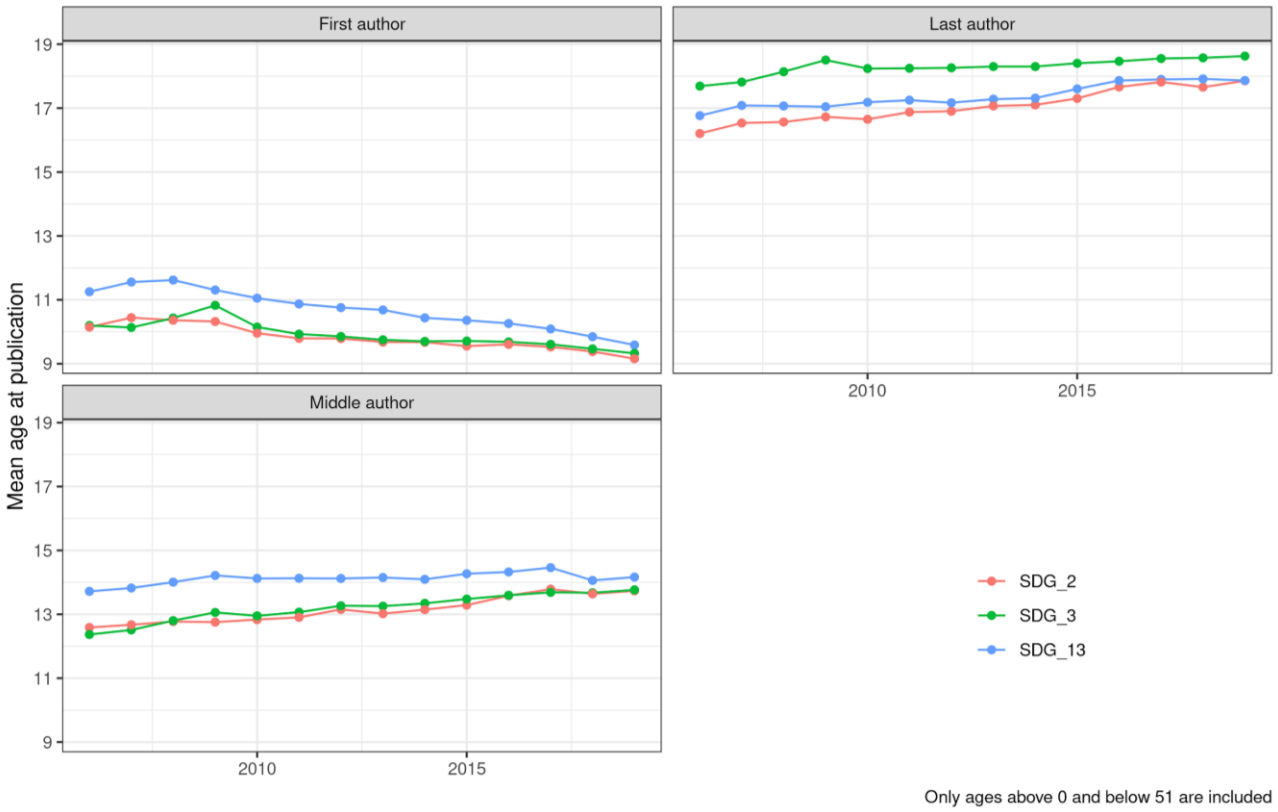

Figure 37: Academic age of authors over time

The gap in ages between first and last authors (Figure 38) is biggest in SDG Health/Well-Being (SDG 3) (7.5 years in 2006) and smallest in SDG Climate Action (SDG 13) (5.5 years in 2006). Since 2006, the gap in age has increased, from 5.5 years in SDG Climate Action (SDG 13) in 2006 to 8.3 years in 2019, and from 7.5 in SDG Health/Well-Being (SDG 13) in 2006 to 9.3 in 2019. The differences between SDGs in terms of this gap are decreasing, however.

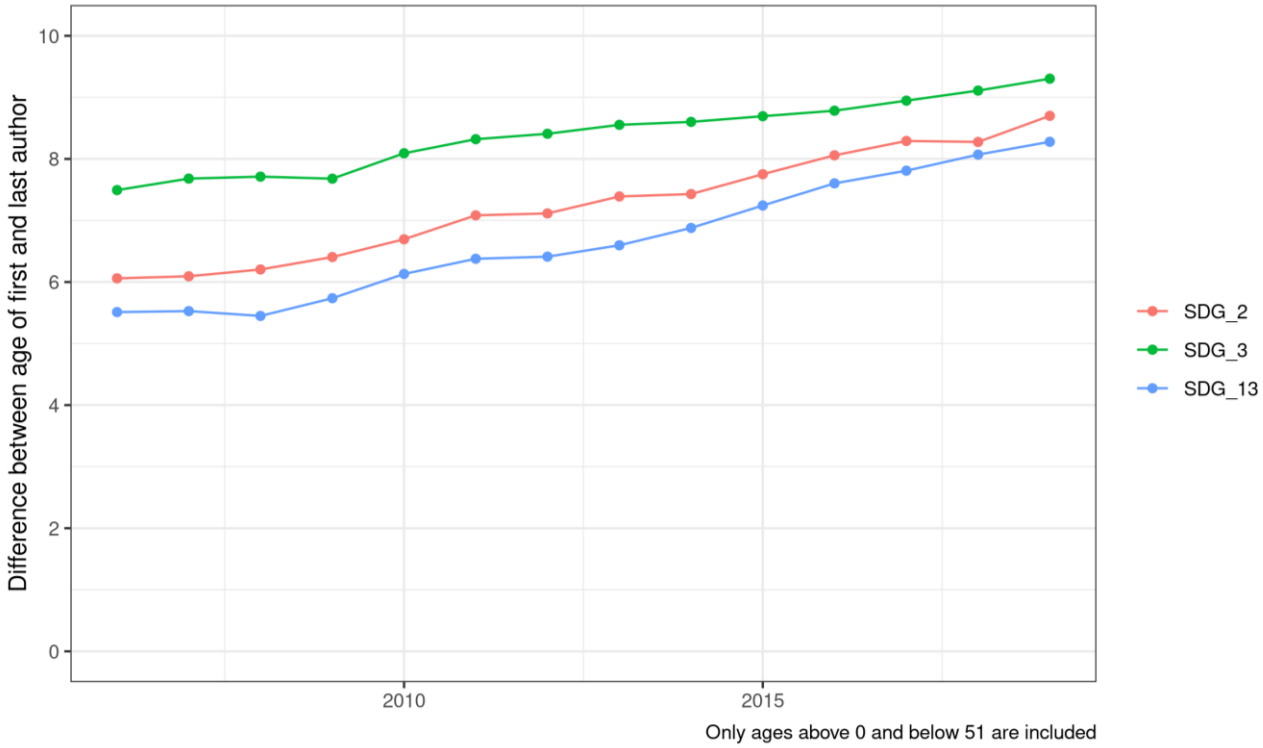

Figure 38: Developing gap in academic ages of first and last authors

Gender

Using the methodology described above (section 5.3.2), we analysed the gender distribution among authors across the three SDGs. We follow West et al. (2013) in investigating the gender division in authorships, i.e. "An *instance of authorship* consists of a person and a paper for which the person is designated as a co-author." (West et al. 2013, 3) Overall, we find the share of female authorships to be lower than that of male authorships (Figure 39). The share of female authorships increases over time across all SDGs. It is highest in SDG Health/Well-Being (SDG 3) (30% in 2006, 37% in 2019), followed by SDG Zero Hunger (SDG 2) (28% in 2006, 35% in 2019), and lowest in SDG Climate Action (SDG 13) (19% in 2006, 27% in 2019).

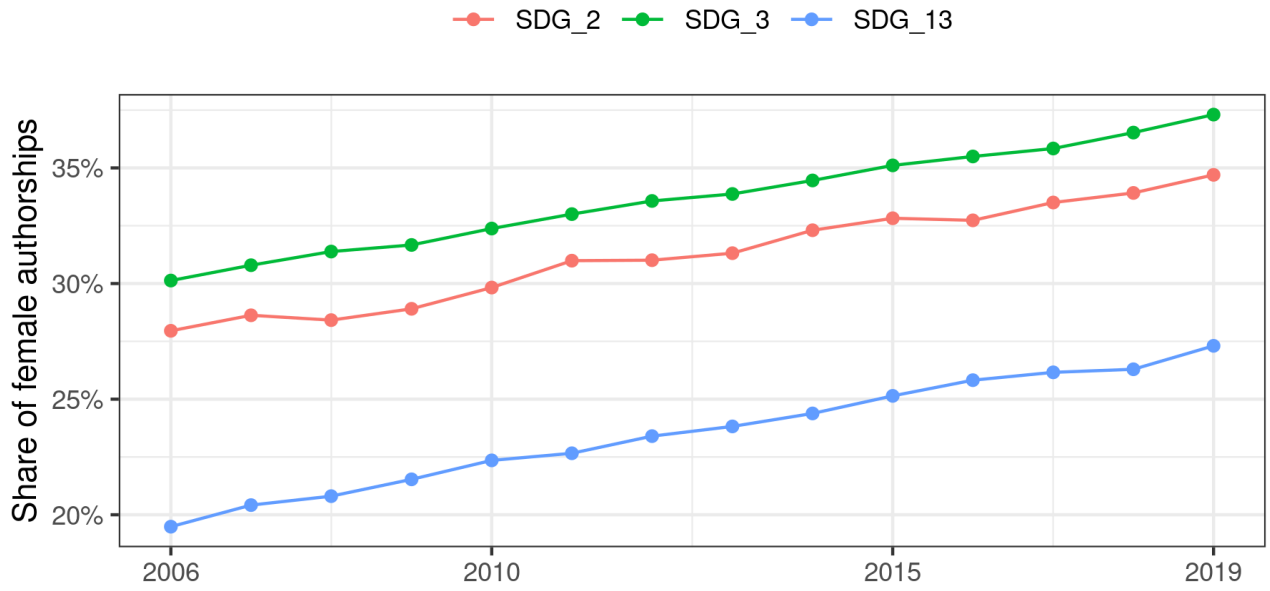

Figure 39: Share of female authorships by SDG

Analysing gender share by author position, we find a clear difference in the gender share of first and/or single authors to last and middle authors (Figure 40). The share of female authorships is comparatively high for first authors (32%-43% in 2019, across SDGs), and low for last authors (22%-30%). The gap in gender shares between SDGs remains stable over time for first authors (about 10-12 percentage points), but increases for last authors, where the share of female authorships grows faster in SDG Health/Well-Being (SDG 3) than in SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13).

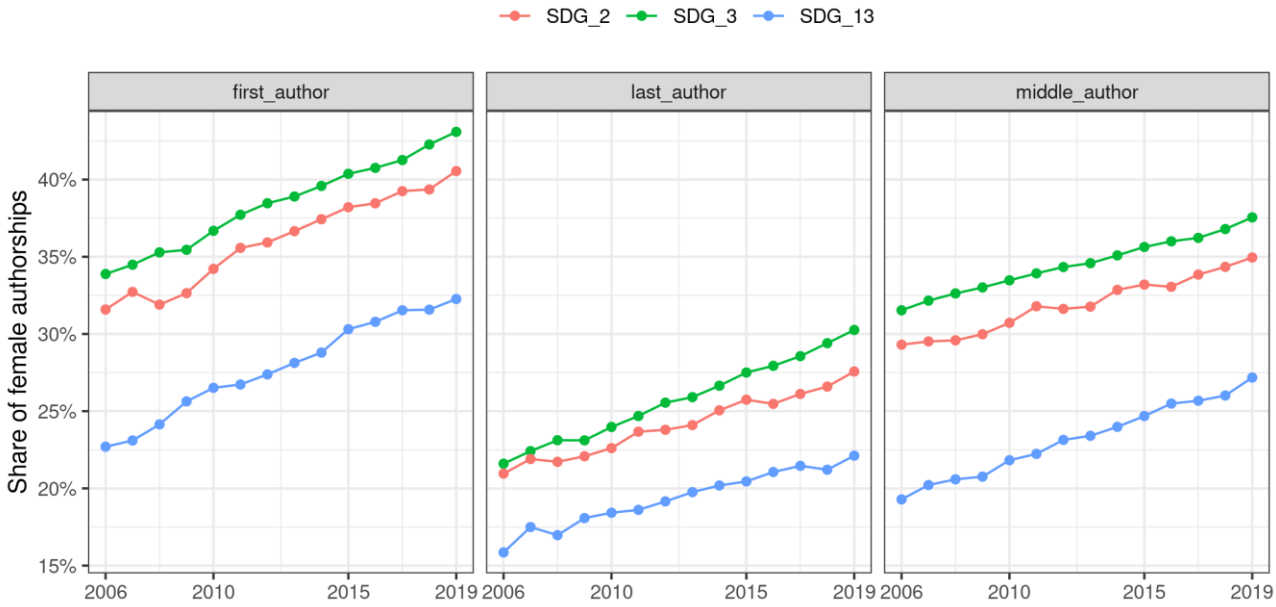

Figure 40: Share of female authorships by SDG and author position

5.5. Authorship positions among publications with international collaboration

Building on the findings regarding general trends in publishing research on SDGs Zero Hunger, Health/Well-Being and Climate Action, we also investigated collaboration. In the following analysis, we only consider publications with at least three authors and with affiliations from at least two distinct countries. The initial hypothesis was that for such publications, authors from lower income countries end up in less prestigious positions on the byline, i.e. middle positions. However, our data does not support this hypothesis. Figure 41 and Figure 42 show the share of authorship positions by continent and income group, stratified by SDG and author position. Overall, we found only minor differences between continents and income groups, with no systematic pattern that is similar across all SDGs.

Regarding the initial hypothesis, the group of low income countries actually has the lowest share of middle author positions in SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13). While this income group has the lowest shares for last authors in SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3) (supporting the hypothesis), it has the highest share of last authors in SDG Climate Action (SDG 13) (opposing the hypothesis). A potential confounding factor to the analysis are differing conventions that vary by country and field, such as ordering authors alphabetically, having PhD students or rather more senior authors in the first position, and so on. Given the macroscopic view we took for this analysis, we are unable to control for such

effects. However, the fact that the findings are consistent across SDGs should be taken as encouraging regarding equity in distribution of credit within international collaborations.

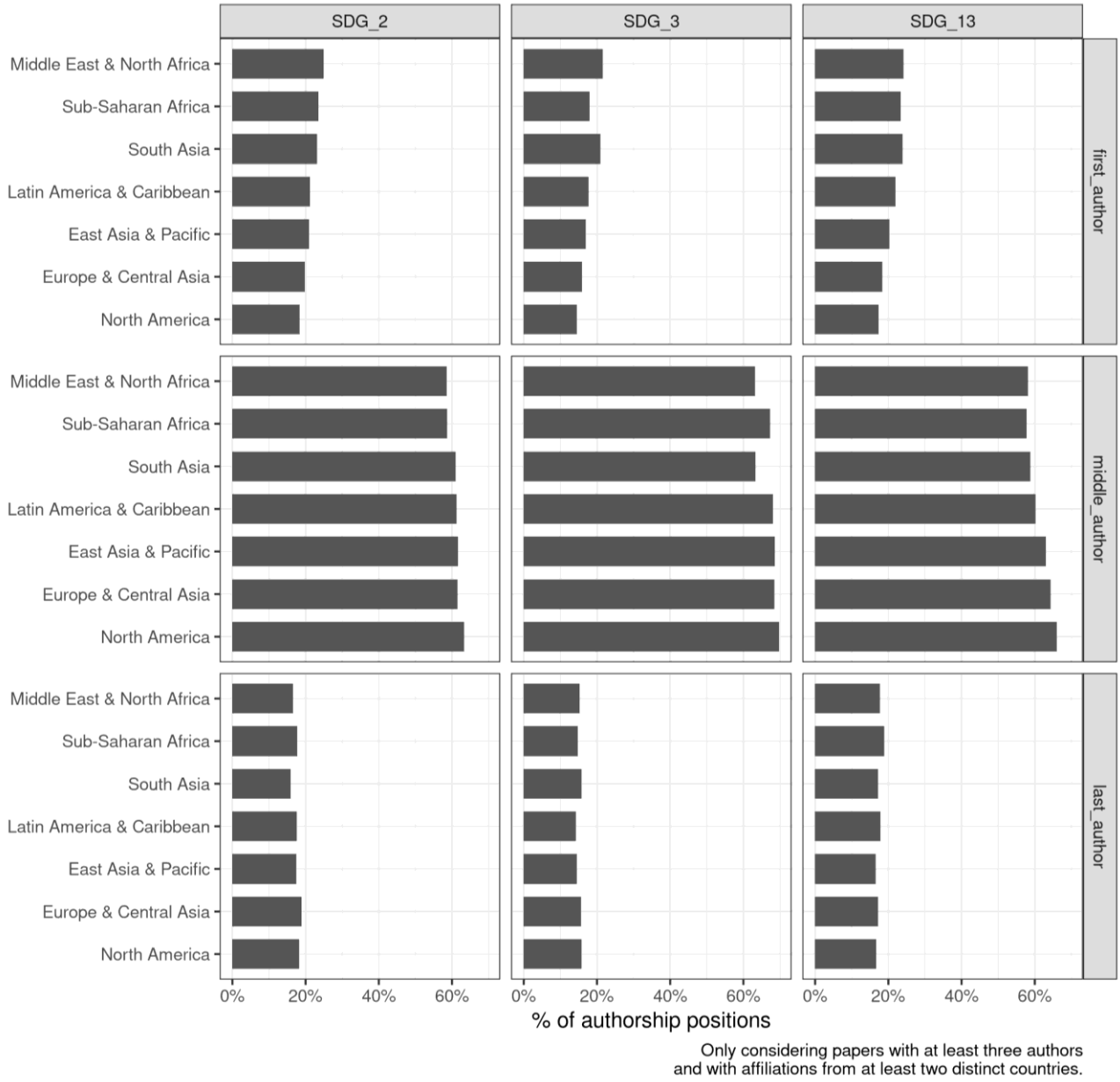

Figure 41: Distribution of authorship positions by world region

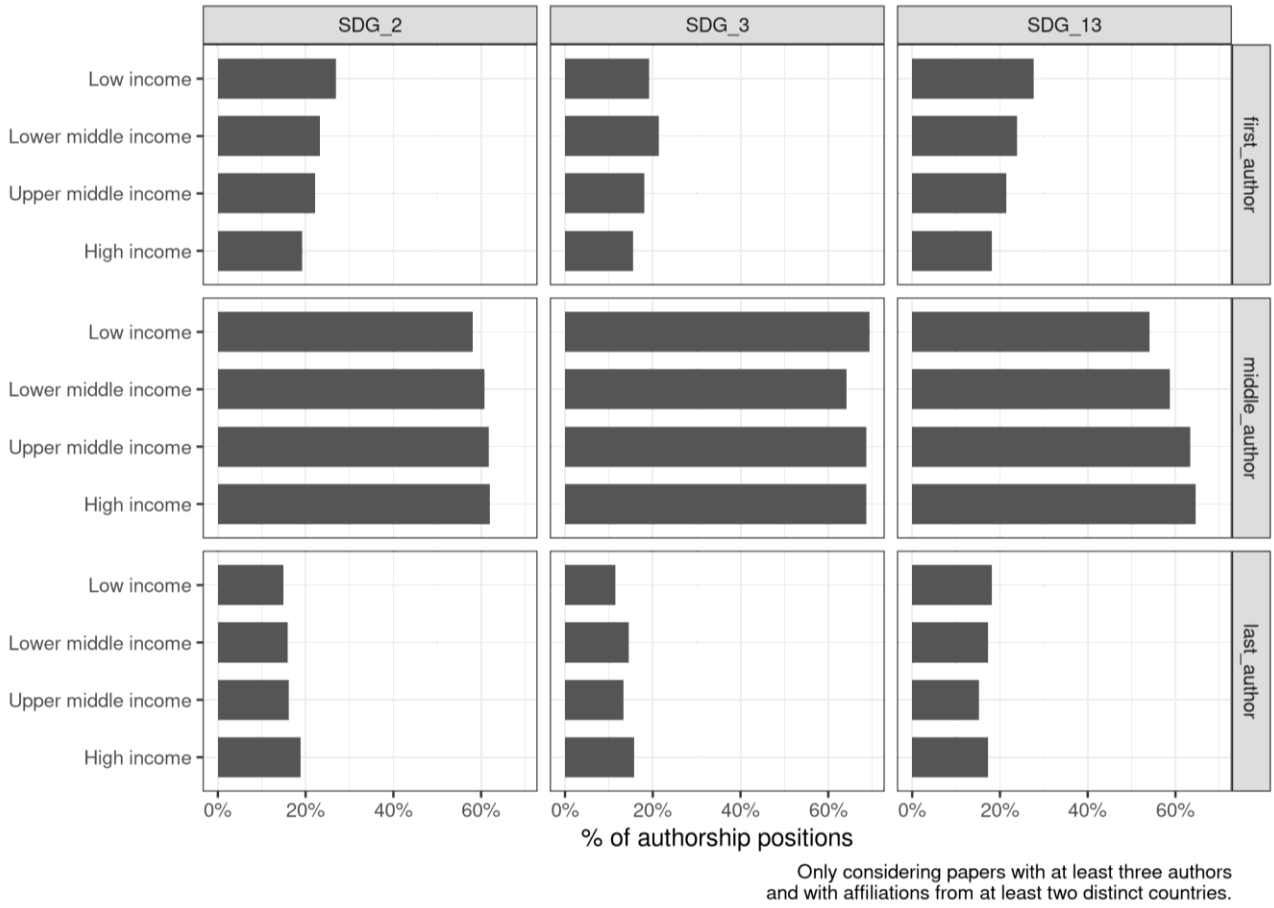

Figure 42: Distribution of authorship positions by income group

5.6. Discussion

In this study we laid the groundwork and established benchmarks for our final study on potential effects of stratification in OA publishing in SDG research. We analysed general properties of our dataset, in terms of its size, and in terms of its distribution along factors such as gender, academic age or institutional prestige.

The amount of research published on the three SDGs grew substantially over the years 2006 to 2019. While the literature reports an average yearly growth rate of about 3% (Ware and Mabe 2015, 28), growth rates in our sample are much higher, ranging from 5-7% on average. In contrast, the impact of individual publications within our sample (as measured by the total number of normalised citations) is declining. If taken at face value, this would be a worrying trend, since it may mean that research tackling key issues related to three UN SDGs is increasingly having a smaller impact in their respective fields (it might also however simply reflect increasing specialisation in a growing area of interest). However, we cannot discount the possibility that the results are driven by particularities of the normalisation procedure used. To increase confidence in these findings, further analysis using alternate modes of normalisation (such as field normalisation (Waltman and van Eck 2012) or more complex citing-side normalisation (Waltman 2016)) should be undertaken.

In terms of the share of female authorships in research on the SDGs, we see an upwards trend, with the share increasing from 19-30% in 2006 to 27-37% in 2019. We found the share of female authorships to be highest in research on SDG Health/Well-Being (SDG 3), and lowest among research on SDG Climate Action (SDG 13).

These findings are similar to previous results from the literature (Larivière et al. 2013; West et al. 2013). Also in line with previous findings, we find a substantial difference in the share of female authorships in terms of authorship positions. The last position on the byline, in many fields in the medical and natural sciences associated with prestige or seniority (Helgesson and Eriksson 2019), has a lower share of female authorships than middle authors, while single and first author positions have a higher share of female authorships. There are no major signs of this gap closing.

In terms of academic age, we similarly find a clear differentiation in terms of authorship roles. Younger researchers are more frequently found in first author positions, while older researchers are more frequently found in last author positions. The gap between the average ages of first and last authors is increasing over the years from 2006-2019, from 5.5-7.5 years in 2006 to 8.2-9.3 years in 2019. A potential explanation for this increase can be found in the continued move towards team science, with increasingly differentiated roles (Fortunato et al. 2018).

Finally, we find that overall institutional prestige is strongly linked to research production in the three SDG areas. Institutions which are included in the Leiden Ranking produce more publications and receive more citations than those not listed in the Leiden Ranking. Within the ranking, higher-prestige institutions again produce more and receive more citations than lower-prestige institutions. The concentration is highest in SDG Health/Well-Being (SDG 3), where the top 20% of institutions produce more than 50% of fractionalised publications, while the bottom 20% of institutions produce slightly more than 5% of fractionalised publications. In turn, the concentration is lowest in SDG Zero Hunger (SDG 2), where the top 20% account for about 40%, and the bottom 20% for slightly less than 10% of fractionalised publications. These shares are stable over time, which indicates persistent stratification within the scientific communities contributing to SDGs Zero Hunger, Health/Well-Being, and Climate Action.

6. Patterns of Stratification in Open Access Publishing across Key UN Sustainable Development Goals

6.1. Introduction

Science can be understood as a game of recognition (Zuckerman 1988). Scientists go about their business, trying to solve puzzles (Kuhn 2012) and seek to be recognized by fellow scientists, through promotions, prizes, and citations (c.f. the concept of the credibility cycle (Latour and Woolgar 1986)). These cycles of reciprocal critique and appraisal of previous research are functional to the scientific enterprise, which builds on previous efforts. This inherent logic of rewarding scientists based on their work's merit, and the subsequent pursuit of rewards and recognition leads to various forms of cumulative advantages (Zuckerman 1988; Merton 1968). For example, if rewards are distributed based on merit, those that received rewards early on will have more resources available to conduct further research. This leads to cumulative advantage over time at the level of institutions, but also at the level of individual careers.

The system of scholarly communication which has evolved based on these basic principles has been found to disadvantage researchers from the Global South, primarily due to high subscription costs for the most prestigious journals (Matheka et al. 2014), but also due to struggles as non-native speakers of English (Ramírez-Castañeda 2020). New developments in the sphere of Open Science aim at democratizing knowledge, creating a more equitable and diverse environment (Fecher and Friesike 2014; Tennant et al. 2016; Ross-Hellauer et al. 2021). As part of this transition to Open Science, Open Access (OA) publishing allows more people than previously to access scientific knowledge. While the increased access to research results is an important step in democratising scientific knowledge, this might shift demands from consumers to producers of scientific knowledge, with high Article Processing Charges (APCs) potentially creating a new source of inequality where researchers with less resources are locked out of the publishing process.

In this study we investigate the potential emergence of new publishing hierarchies. Does the uptake of OA publishing change existing hierarchies within academic publishing, and if so, in which ways?

We focus on the question of who publishes where. Previous studies (e.g., Siler et al. 2018; Olejniczak and Wilson 2020) suggest an emerging division in OA publishing related to individual and institutional resources. Our research builds on these previous efforts and incorporates the global dimension: are authors from richer countries more likely to publish OA in research on key UN SDGs, and especially via options involving an APC? Is the association between individual level attributes such as academic age and gender and the likelihood of publishing OA similar across strata of institutional prestige and across countries?

6.2. Background

6.2.1. The emergence of OA publishing - trading equality of access with increased inequality in production of knowledge?

The growth of OA publishing has been well documented. Piwowar et al. (2018) conducted a landmark investigation into the prevalence, growth and impact of OA articles. They found the share of research articles being available as OA increasing quickly, with an estimated 44.7% of all articles being OA in 2015. Their study also corroborates the OA citation advantage, concluding that OA publications received on average 18% more citations than closed access papers. However, growing evidence suggests that current OA publishing models tend to advantage researchers endowed with more resources (Ross-Hellauer et al. 2021; T. van Leeuwen 2019). These advantages operate on multiple levels: that of demographic characteristics of authors, the level of institutions, and the level of countries or world regions.

Focusing on the individual level, Olejniczak and Wilson (2020) investigated which author characteristics are related to OA publishing. Overall, they found a higher likelihood for publishing OA articles that involve an APC for authors of male gender, from prestigious institutions, with previous federal (US) research funding, or an association with a STEM field. They conclude that "*[p]articipation in APC OA publishing appears to be skewed toward scholars with greater access to resources and job security.*" The role of institutional support in covering APCs is evidently of urgency for researchers without an affiliation to a research-oriented institution. High APCs and APCs in general might preclude this growing segment of researchers from contributing to the scientific record (Gray 2020, 1673; Burchardt 2014; ElSabry 2017).

Regarding institutional characteristics, Siler et al. (2018) found a clear hierarchy in publishing access outcomes. Analysing a set of articles from health research, they found that authors from lower-ranked institutions were more likely to publish in toll-access journals, as well as in OA journals with no article processing charge (APC). Regarding the average APCs paid by the different types of institutions, authors affiliated with higher-ranked universities, hospitals and non-profit organisations generally paid higher APCs than authors affiliated with companies, governments, research institutions, scientific associations as well as lower-ranked universities.

These differences on individual and institutional levels are complemented by inequalities of access to scientific publishing on the level of countries and regions. Researchers from the Global South have more difficulties in paying increasingly high APCs simply due to lower purchasing power (Demeter and Istratii 2020; Matheka et al. 2014). Waivers for APCs do exist, but are not always effective in countering this issue (Ross-Hellauer et al. 2021). The discrepancy in access to publishing is linked to the broader system of knowledge production and its global distribution. Research from the Global North is often self-centred, with a focus on phenomena and viewpoints which are relevant to those countries (Czerniewicz 2015; Collyer 2018). This is reflected in which types of research are accepted in the most prestigious journals. In the Global South on the other hand, there is a strong focus on publishing in these prestigious journals. Publications in highly prestigious journals are sometimes rewarded directly in terms of cash payments (although this practice was abolished in 2020 in China (Mallapaty 2020)), but also indirectly through higher chances to receive promotions. This leads to a situation where for researchers from the Global South to publish in highly regarded journals, they not only have to align their research with that of the North's agenda, but also to pay even higher levels of APCs, since perceived journal prestige (represented by common measures such as the Impact Factor or the DOAJ SEAL) and APC prices are closely linked (Demeter and Istratii 2020; Gray 2020;

Siler and Frenken 2019). In economic terms, research money from low-income countries (LIC) partly subsidises research from the North, with researchers from less industrialised countries publishing considerably more frequently in mega-journals such as PLOS ONE than in the publisher's more prestigious counterparts like PLOS Biology (Ellers, Crowther, and Harvey 2017).

Finally, these tendencies might lead research published in local journals to become less visible. As high-income countries (HIC) enforce policies to publish OA, research from LIC which might not yet be OA becomes even less visible (Czerniewicz 2015). Since local journals also usually have lower rankings on common metrics such as the journal impact factor, research published in these journals not only receives less exposure, but might be perceived as to be of lesser quality (Gray 2020).

6.2.2. Country differences in the production of OA publications

To date, not much research has been undertaken into how OA production rates differ between countries (but see e.g., Huang et al. 2020; Robinson-Garcia, Costas, and Leeuwen 2020). Investigating publications from the biomedical sciences, Iyandemye and Thomas (2019) found a high share of OA production in low income countries, particularly by researchers from sub-Saharan Africa. Another factor associated with higher rates of OA publication was found to be publications written with international collaboration. Iyandemye and Thomas (2019) conclude that OA policy and OA publication rates do not align well, on the global scale.

6.2.3. Funder and country mandates

A strong driver in the emergence and uptake of OA is funder mandates. Lariviere and Sugimoto (2018) found about two-thirds of research funded by prestigious funders like the NIH, NSF, Wellcome Trust, or the ERC to be available as OA. There was high variance between funders, however, with high compliance among research funded by the Wellcome Trust, NIH, or the Medical Research Council (UK), but low compliance among Canadian research funders and the NSF. Potential reasons for low compliance were posited, in that some policies mandated "voluntary compliance" (NSF) or allowed authors to deposit papers after publication (Canadian research councils). Huang et al. (2020) investigated a set of top-performing universities from across the world and found increased OA uptake after policy changes that e.g. tie funding to OA publishing (the Research Excellence Framework (REF) in the UK, or funding mandates from the EC (Athena Research & Innovation Center et al. 2021)) or resulted in "transformative agreements"[35] with major publishers (e.g., deals with Springer and Wiley in the Netherlands).

6.3. Data & Methods

6.3.1. APC data

In this study, we will not only analyse publications with regard to their OA status, but also in relation to the journals' policies in terms of APCs. To facilitate this analysis, we use data from the Directory of Open Access journals (DOAJ). We downloaded the full DOAJ dataset on 2021-07-19, and merged it with the list of journals available in MAG. Since data from DOAJ includes two ISSNs (print ISSN and EISSN), but MAG only includes a single ISSN, we took multiple steps. We first matched journals from DOAJ to MAG via the EISSN. In the second step, we matched the remaining journals via the print ISSN. Finally, we matched the remaining journals

[35] In transformative agreements, institutions like universities or libraries remunerate publishers for the costs associated with OA publishing (see e.g., Borrego, Anglada, and Abadal 2021).

directly by name (exact strings). Overall, we were able to match 6,980 journals out of 16,623 journals from DOAJ to MAG.

To facilitate the comparison of APCs, we always used the USD APC values, if available, and converted all remaining APCs into USD with the exchange rate of 2021-07-19 (using the *getFX* function from the quantmod package (Ryan and Ulrich 2020), following Gray (2020)).

6.4. Results

6.4.1. Availability of SDG research as OA

As already stated, previous research has found a strong growth in the proportion of research available as Open Access. Among the publications in our sample, we find a similar increase, from below 30% OA in 2006 to above 50% OA in 2018 (Figure 43). Research in SDG Health/Well-Being (SDG 3) is consistently above SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13) in terms of OA availability, however.

Figure 43: OA share by SDG

Breaking down OA publishing into publisher- and repository-driven OA[36], we find a clear trend towards publications being made available through journals and repositories at the same time (e.g., depositing a preprint in a repository and then publishing in an OA journal) (Figure 44). The share of publications that are OA both through repositories and journals increased from about 30% in SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3) in 2006 to about 50% in 2018. OA availability through journals and repositories

[36] See section 2.5 on how we operationalised these terms.

simultaneously of research on SDG Climate Action (SDG 13) also increased, but from a higher initial level (just below 40%).

The share of articles where OA is only provided through repositories decreased, from above 30% of all publications in 2006 to below 20% in 2018. The share of publications that are OA solely through OA journals first decreased from 2006, and has seen a slight increase since. This initial decrease and subsequent increase is driven mostly by SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3), with research on SDG Climate Action (SDG 13) exhibiting a slight upward trend throughout.

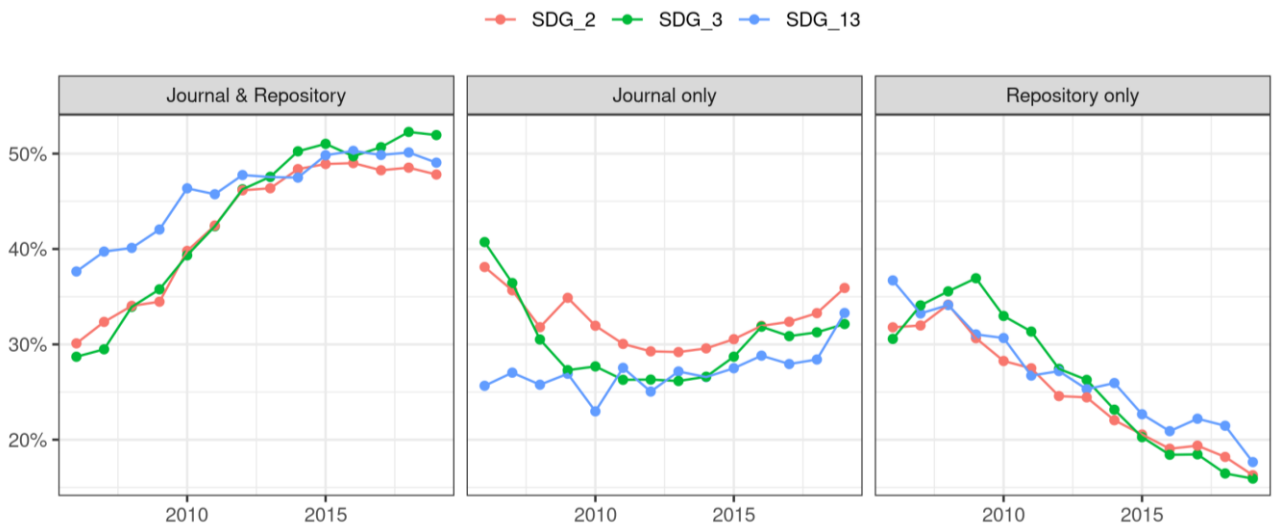

Figure 44: OA share by SDG and hosting type

Academic age

Considering academic age, we investigated whether age was associated with OA publishing (Figure 45) We consistently find that articles with junior researchers in first, last and middle positions on the byline are less frequently available as OA than research published by older researchers. We assume that first and last authors have more influence on publishing decisions, and subsequently analyse patterns for them.

Starting with first authors, we find that the age of researchers has the smallest association with OA publishing in SDG Health/Well-Being (SDG 3), since differences in the share of OA publications are only minor (below 5%). The association is stronger in SDG Zero Hunger (SDG 2), and strongest in SDG Climate Action (SDG 13), where the difference in OA publication shares for 2019 between the most junior and most senior researchers is 18.5%.

Considering last authors, differences between junior and senior researchers are slightly higher for SDG Health/Well-Being (SDG 3) compared to first authors, but lower for SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13). Still, SDG Climate Action has the highest distance between the most junior and most senior group (12.7%), compared to a difference of 6.6% for SDG Health/Well-Being and 5.4% for SDG Zero Hunger in 2019.

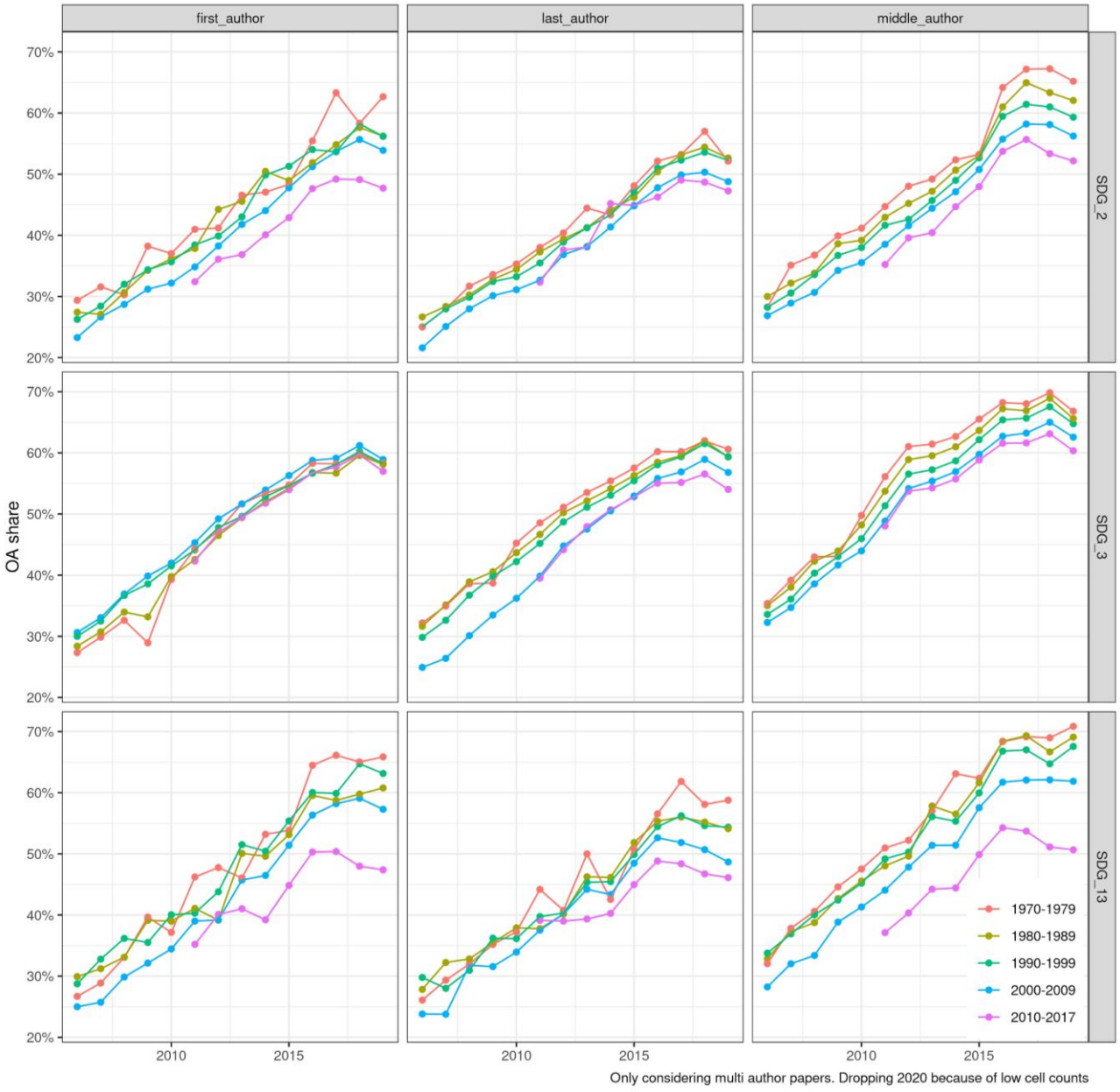

Figure 45: OA shares by academic age over time. Seniority ranges refer to the first year publishing a scholarly paper. We only consider publications with at least three authors. Numbers for 2020 were removed due to low cell counts.

Gender

Moving on to author gender, we investigate whether gender is associated with higher or lower OA publishing propensity. We calculated the share of female authorships separately for OA and non-OA publications (following the same method as in section 5.4.1). Here, we find a small but consistent higher share of female authorships among OA articles compared to non-OA articles in SDG Health/Well-Being (SDG 3) (Figure 46). The picture is less consistent for SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13), where in some years there is a higher share of female authorships among OA publications, and in some years among non-OA publications. For the most recent years (2016-2019), we observe a consistent increase of the overall share of female authorships among OA publications but a slight decline among non-OA publications for SDG Zero Hunger and SDG Climate Action.

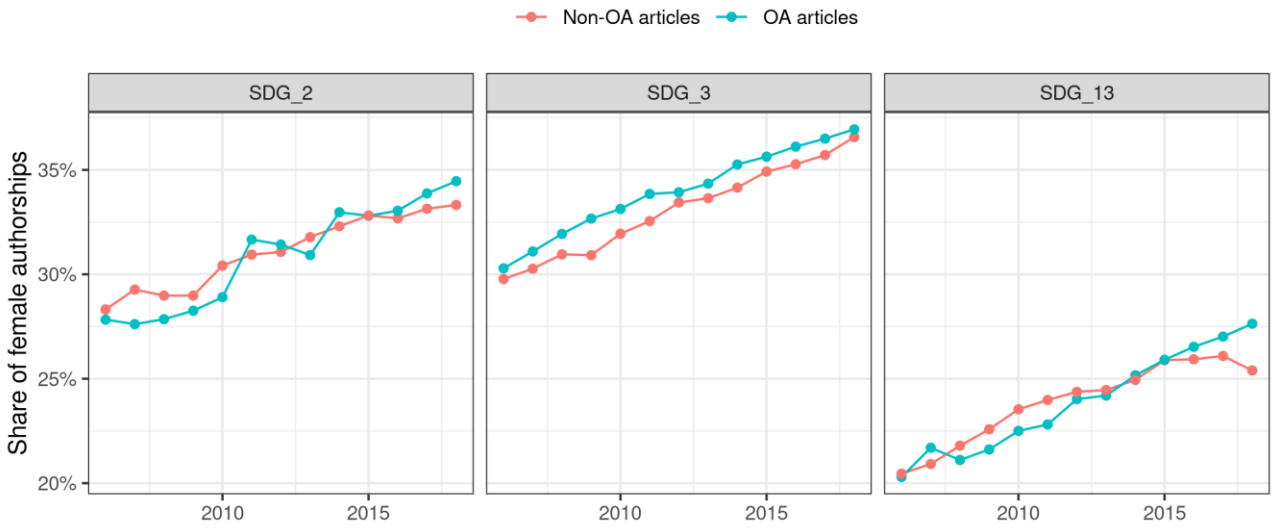

Figure 46: Share of female authorships by OA status

Turning to authorship positions, we use a slightly different approach. First, we select only papers with more than one author to be able to differentiate between first and last author positions. We then calculate the share of publications that are OA, stratified by author position, SDG and gender. Last, we calculate the difference of OA publishing propensity by subtracting the value for females from the value of males. Figure 47 displays this difference: positive values indicate more OA publishing among females, negative values indicate more OA publishing among men.

Overall, we find a slight upward trend, indicating that over time, more research authored by women is being published OA than research by men (Figure 47) First authors in SDG Health/Well-Being (SDG 3) are an exception to this general trend. While for last authors, and first authors from SDG Zero Hunger (SDG 2) SDG Climate Action (SDG 13), men published OA more frequently before 2010, female first authors in SDG Health/Well-Being published more OA in this earlier period, with a slight decline since. Note that these differences are relatively small, with substantial variability between years. Nevertheless, these findings are in line with the above analysis, thus validating the different approaches against each other.

Figure 47: Difference in OA publishing shares between genders over time

Institutional prestige

Considering institutional prestige next, we analysed its association with OA publishing. Higher ranked institutions consistently publish OA more frequently than lower ranked institutions in SDG Health/Well-Being (SDG 3) (Figure 48). The general pattern also holds for SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13), albeit the ordinal order does not always hold there, i.e. sometimes the first quartile (p[0,25]) publishes more OA than the second quartile (p[25,50]). As Figure 49 shows, this association is on a similarly high level for SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3), but substantially lower for SDG Climate Action (SDG 13). Furthermore, for SDGs Zero Hunger and Health/Well-Being, the association weakened over time. Institutional prestige thus has a lower association with publishing outcomes in 2018 compared to 2008.

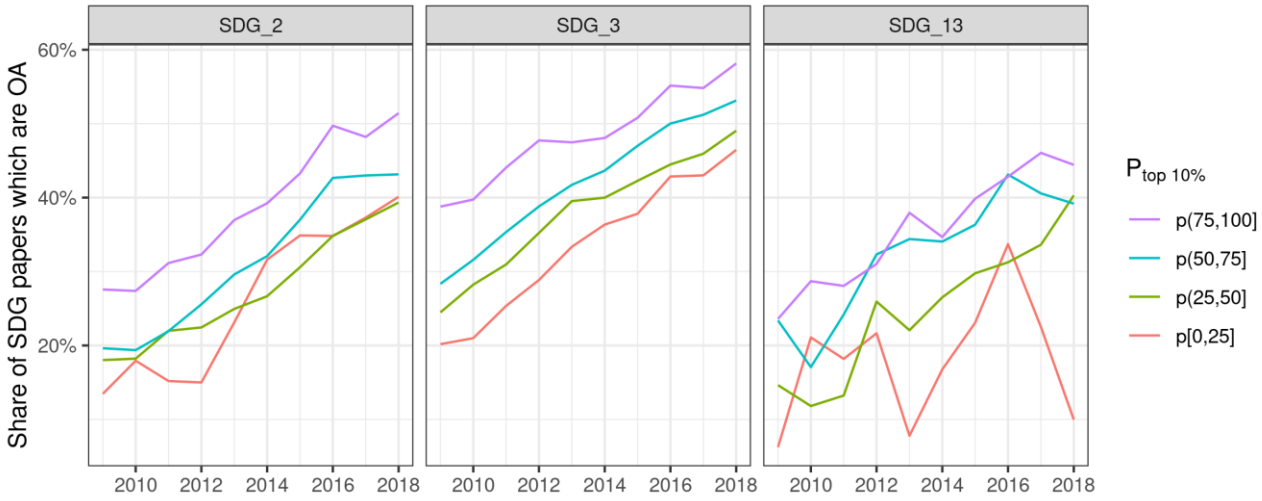

Figure 48: OA shares by institutional prestige

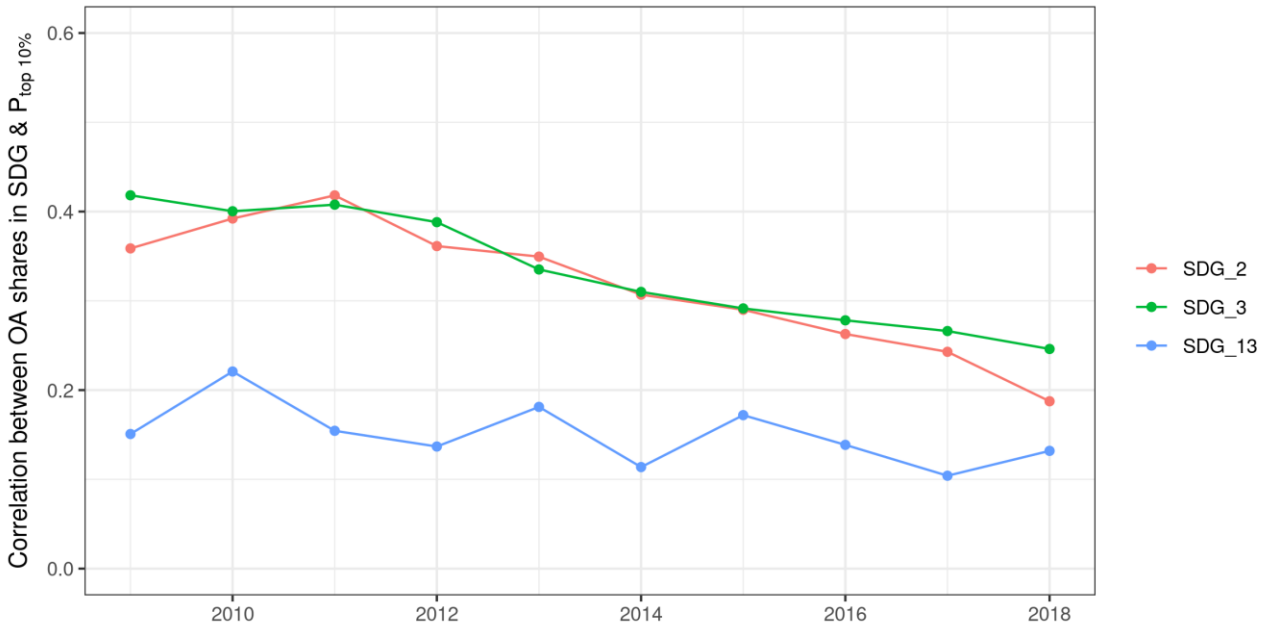

Figure 49: Correlation between OA shares and institutional prestige

As indicated in the previous Study (section 5), the share of OA publications jointly hosted by journal and repository is higher than content that is either journal or repository hosted, and has increased over the last two decades. Comparing hosting types with institutional prestige, we find weak relationships (Figure 50). Higher ranking institutions publish less OA that is solely repository hosted, but more that is either journal hosted or both journal and repository hosted.

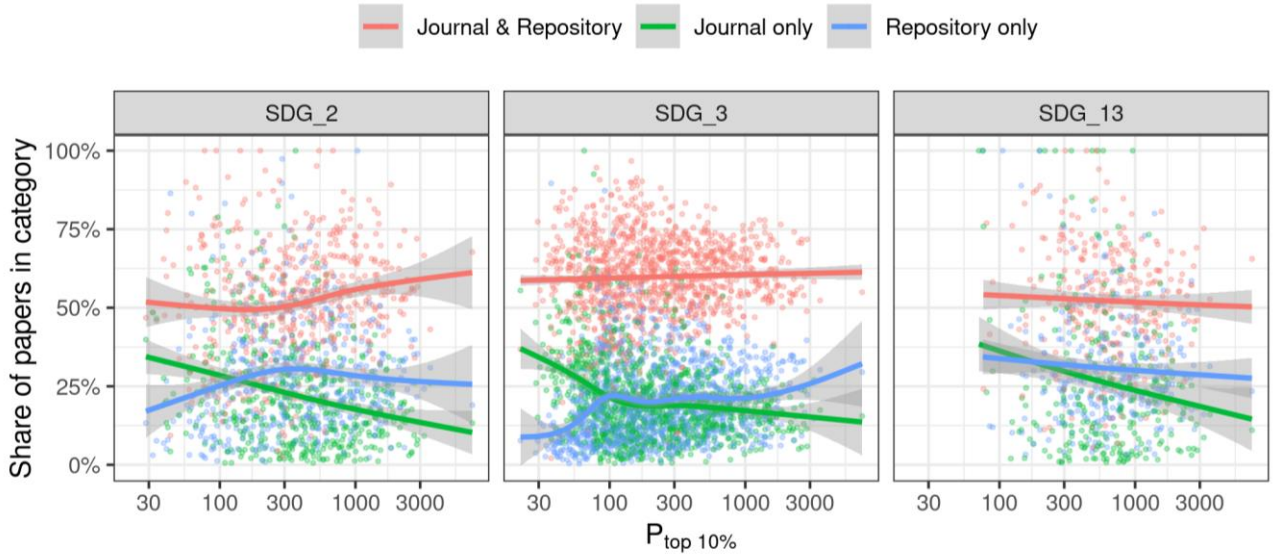

Figure 50: Association between OA hosting type and institutional prestige (2018)

If we consider these relationships over time, we find the correlation between institutional prestige and the number of articles published via the three hosting types to be relatively stable (Figure 51). The negative correlation between ranking position and the share of publications which are hosted solely by journals is weakening slightly. The same holds true for the relationship between ranking position and repository hosted publications, while the correlation between ranking position and the share of publications which are journal

and repository hosted is increasing for SDGs Zero Hunger (SDG 2) and Health/Well-Being (SDG 3), but not for SDG Climate Action (SDG 13).

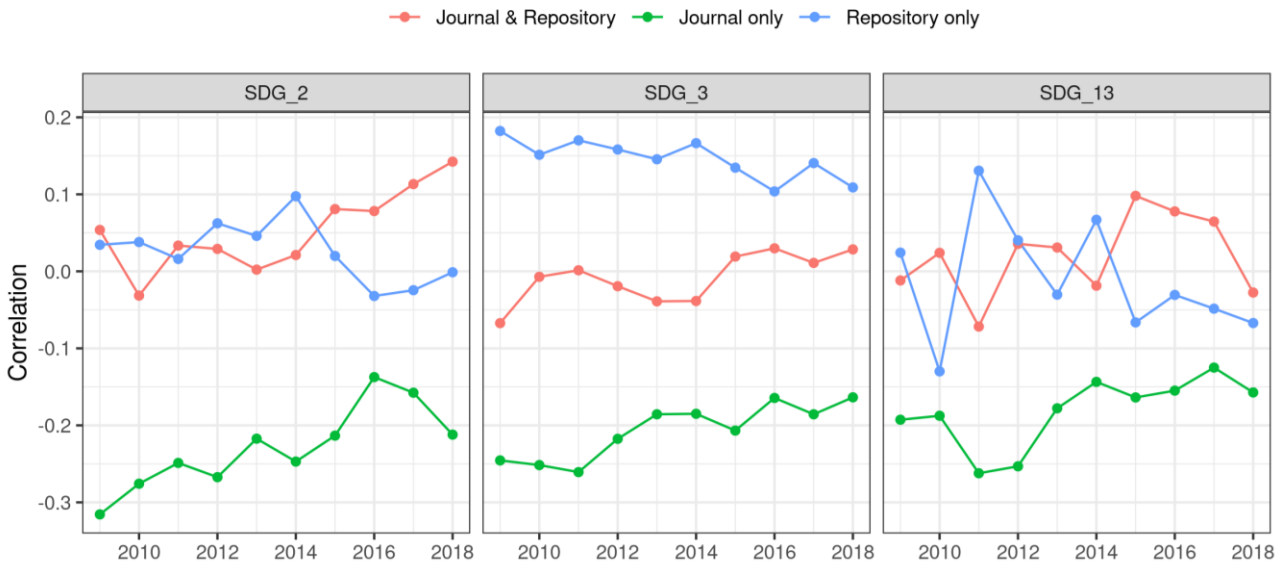

Figure 51: Correlation between the share of publications in hosting types and institutional prestige. Institutional prestige is operationalised with $P_{top\ 10\%}$

Country

Similar to previous findings in the literature (Iyandemye and Thomas 2019), we find a higher share of OA publications among lower income countries (LIC) for publications from 2015-2018 (Figure 52). This association holds for countries up to about 30,000$ per capita. For countries with higher income per capita, the association is reversed, i.e. the highest income countries (HIC) have higher rates of OA publishing than middle income countries (MIC). This effect holds true for all three SDGs in our sample, albeit in slightly different ways (Figure 53). In SDG Climate Action (SDG 13), rates of OA publishing among high and low income countries are similar, with MIC being the lowest. In SDG Health/Well-Being (SDG 3), LIC have higher values of OA publishing than the rest, with MIC and HIC exhibiting similar levels of OA publishing. In SDG Zero Hunger (SDG 2), the lead in OA publishing of LIC is slightly lower, whereas the lead in HIC is larger than for SDG Health/Well-Being (SDG 3).

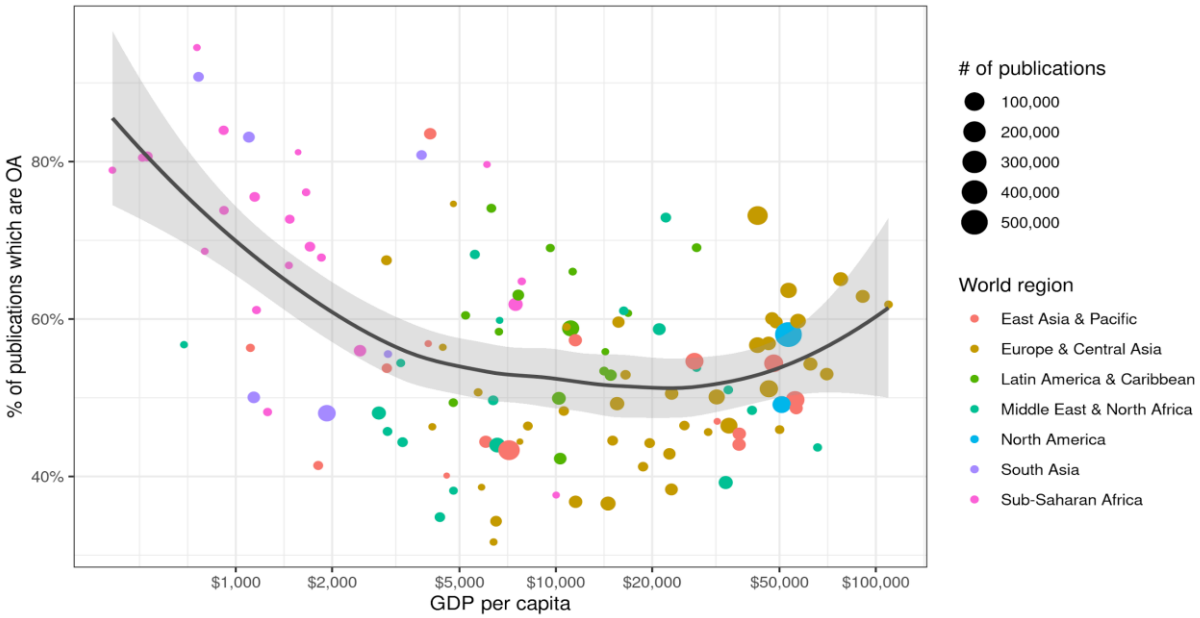

Figure 52: OA publication share and country income. The OA publication share is based on publications from 2015-2018. GDP per capita is the average of 2015-2018. Including countries with 50 or more fractionalised publications in 2015-2018

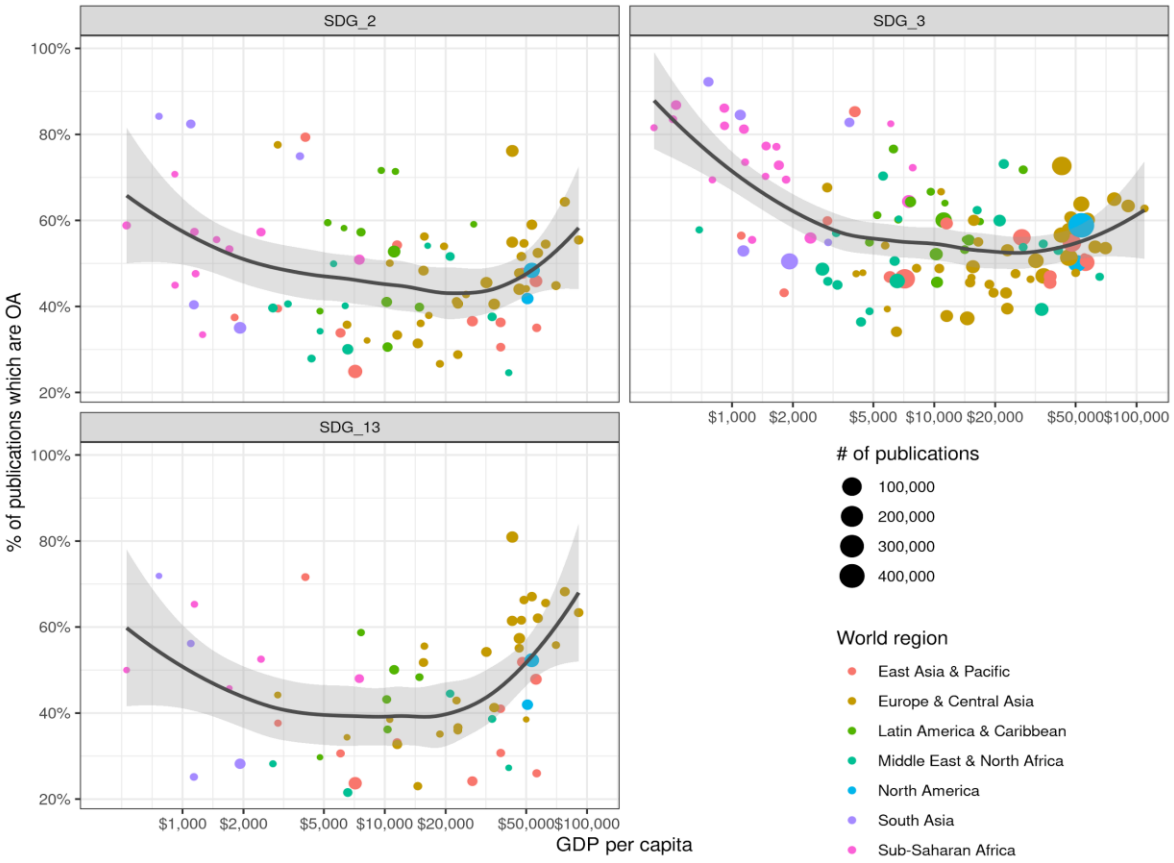

Figure 53: OA publication share and country income by SDG. The OA publication share is based on publications from 2015-2018. GDP per capita is the average of 2015-2018. Including countries with 50 or more fractionalised publications in 2015-2018

We further investigated whether rates of growth of OA publishing differed across countries, continents and income groups. Figure 54 displays the percentage point increase of all countries between two time windows: 2009-2013 and 2014-2018. An increase of 10 percentage points therefore refers to an increase from 30% to 40% OA publishing, or from 70% to 80% OA publishing. Each dot represents one country, and the boxplot is built from the country values. The median, therefore, represents the median of the percentage point increase of the countries' OA shares.

We find the overall percentage point increase to be slightly above 10%. The increase is highest for Latin America & the Caribbean, and lowest for North America (only including the USA and Canada). Sub-Saharan Africa exhibits a high variability, with an almost equal spread between no increases and 30%-point increases. Changes in OA publication shares for Europe & Central Asia are more concentrated, albeit this group also has the highest increase (Ukraine) and the highest decrease (Bosnia and Herzegovina).

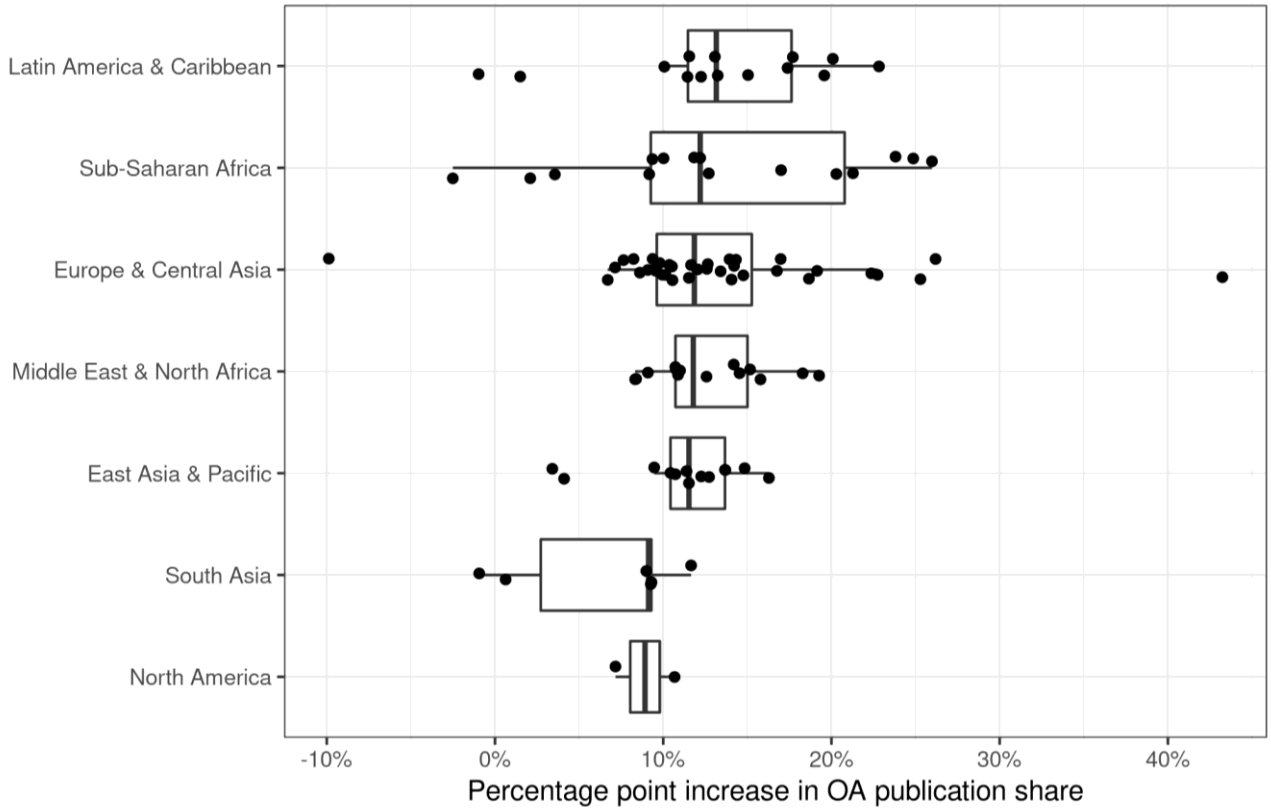

Figure 54: Change in OA publication propensity across regions. Time window: 2009-2013 vs. 2014-2018

Considering income groups instead of world regions, we find lower middle income countries to have the lowest median growth in OA publishing share (Figure 55). Although the median growth of OA publication share is highest for LIC, we only have 5 countries in this group, with high variability.

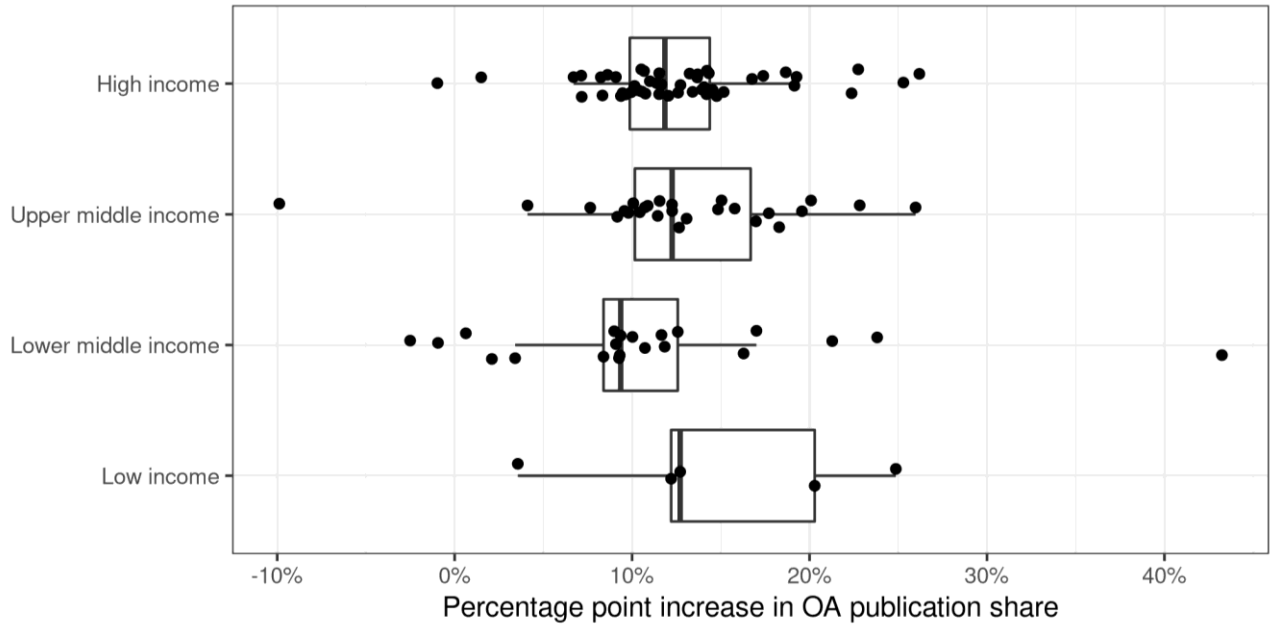

Figure 55: Change in OA publication propensity across income groups. Time window: 2009-2013 vs. 2014-2018

Next, we consider the same question as above, but split by SDG (Figure 56). Here, the individual dots represent countries, i.e. the share of publications that are OA which come from a given country. The black diamonds represent the OA share across all papers of a given region, i.e. not the average of the country means, but the average across all papers. This difference is of importance for continents like North America, where the USA publishes more than Canada, and much more than Mexico. The mean of the countries' shares would be higher, given Mexico's relatively high share of OA publications in SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3), while the mean OA share of all publications from North America is close to the share of OA publications from the USA. Sub-Saharan Africa has the highest OA share in SDG Zero Hunger and SDG Health/Well-Being, but not in SDG Climate Action (SDG 13). Further general trends are similar across all SDGs: Europe, North America and sub-Saharan Africa are generally highest in OA publishing, followed by Latin America. East Asia & Pacific regions, middle east and North Africa, as well as South Asia have the lowest OA rates.

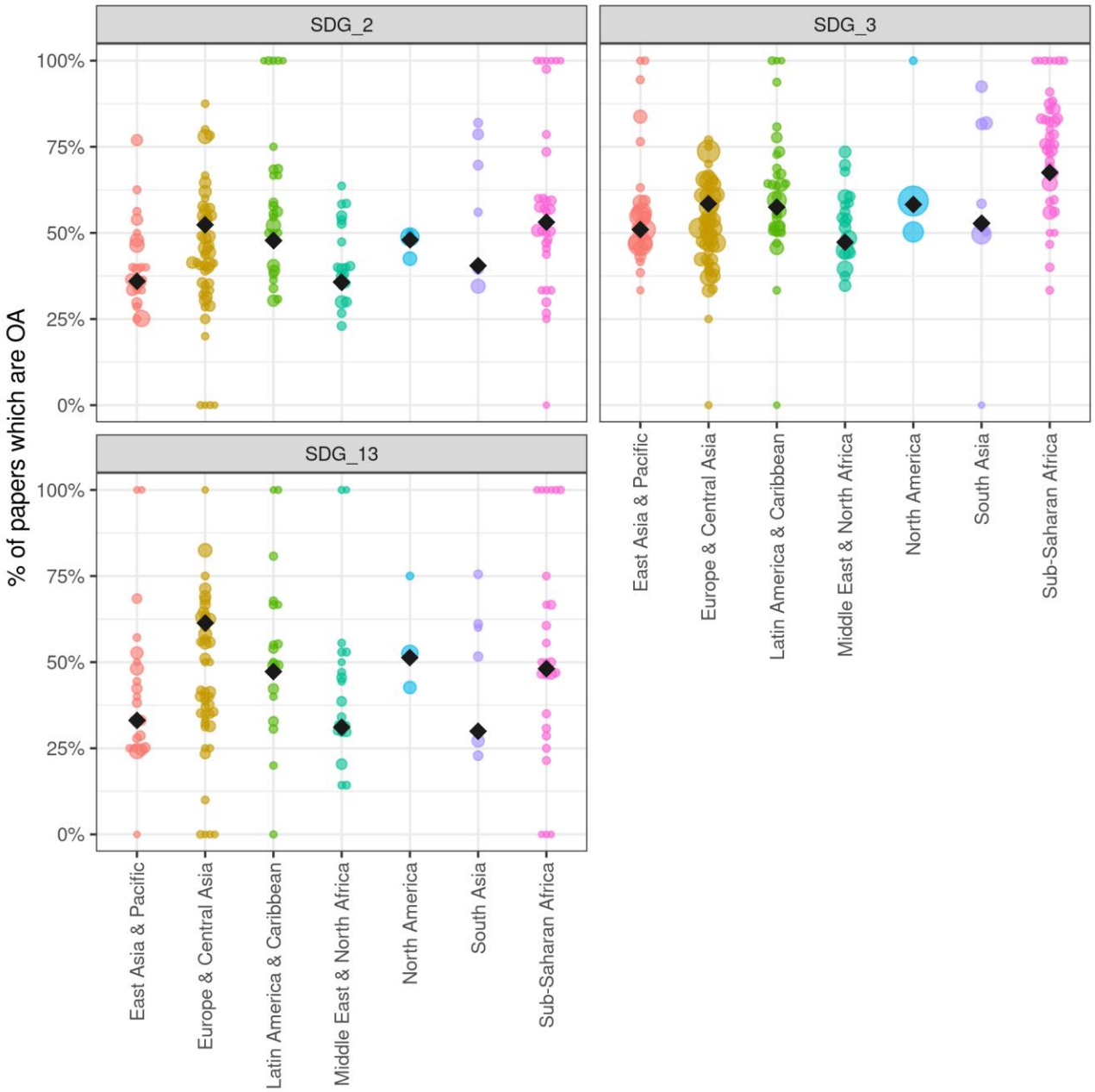

Figure 56: OA publication share by world region. Shares are calculated with full counting, i.e. each authorship counts the same, regardless of the number of total authors on a publication.

Considering income groups, we find similar patterns (Figure 57): LIC have the highest rates of OA publishing in SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3), but not in SDG Climate Action (SDG 13). Upper middle income countries (UMIC) have the lowest rates of OA publishing, with lower middle income countries (LMIC) appearing between UMIC and HIC.

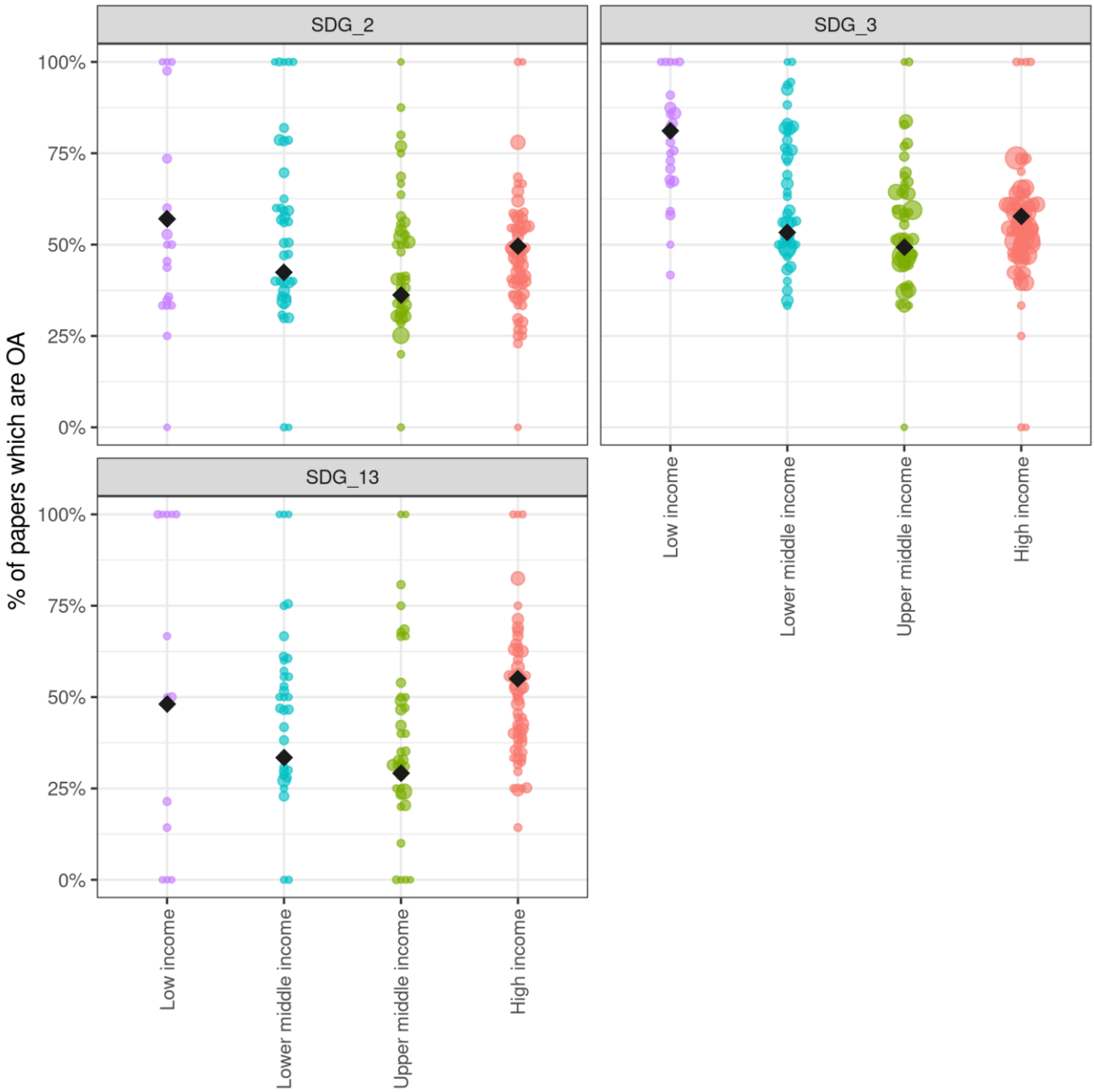

Figure 57: OA publication share by income group. Shares are calculated with full counting

6.4.2. Patterns of stratification: the case of APCs

While there is considerable variety in the uptake of differing OA publishing pathways (through repositories, through journals, or through both), there is also variation within journal hosted OA publishing, since this often entails authors paying an APC. In the following we continue investigating which scholars/institutions tend to publish through which OA pathways by analysing APCs.

After matching articles from our sample to data from Unpaywall (on Methods, see section 2.5), we find that for 54.4% of publications which are OA through journals, the journal is not listed in the DOAJ. Investigating these publications, we find that about 75% of them are either hybrid or bronze OA. The remaining 20-25% are from gold OA articles, which might be from journals which are not in the DOAJ or from journals where the matching from DOAJ to MAG failed. For our further analysis, we only consider publications that are from journals where we have data from DOAJ.

Among the articles matched to journals in the DOAJ, the majority (81.6%) were published in journals that charge an APC. This share is quite similar between SDGs Zero Hunger (SDG 2) and Health/Well-Being (SDG 3) at 81-83%, but considerably higher in SDG Climate Action (SDG 13) with 93.7% of articles being published in a journal charging APCs (Figure 58).

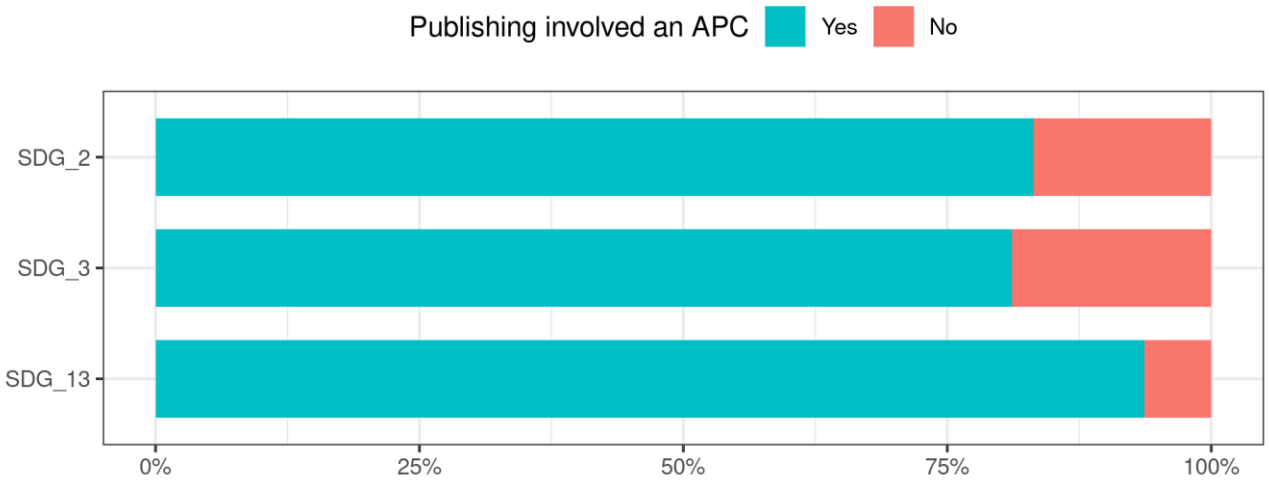

Figure 58: Share of publications involving an APC

This share of APC-based OA does not exhibit a clear trend over the years 2006-2019 (Figure 59). For SDG Zero Hunger (SDG 2), there is a drop in the share of publications to journals involving an APC in the years 2016 and 2017, while there is a similar drop for SDG Health/Well-Being (SDG 3) for the years 2011-2016. We are uncertain about the causes for these declines and recoveries, which will be investigated in future work.

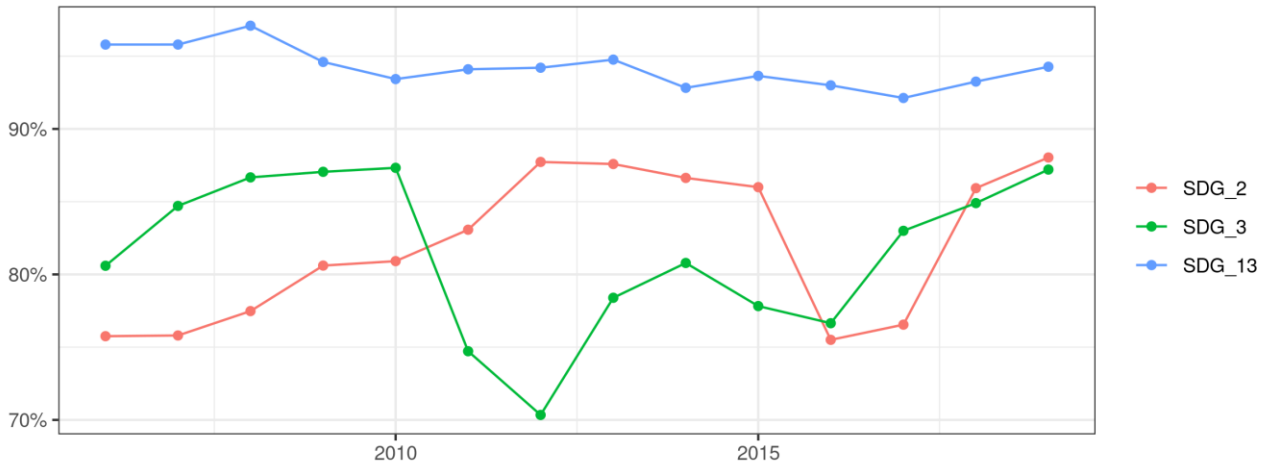

Figure 59: Share of publications from journals that charge an APC

Institutional prestige

We now compare the data on APCs with institutional prestige via the Leiden Ranking. We find that for the period 2015-2018, higher ranking institutions publish more frequently in journals that involve paying an APC (Figure 60). The effect is strongest in SDG Zero Hunger (SDG 2), where the lowest ranking institutions have a markedly lower share of publishing with APC journals than higher ranking journals. SDG Health/Well-Being (SDG 3) exhibits an almost similar effect, whereas in SDG Climate Action (SDG 13) the ranking has a very weak association with publishing outcomes in terms of APCs.

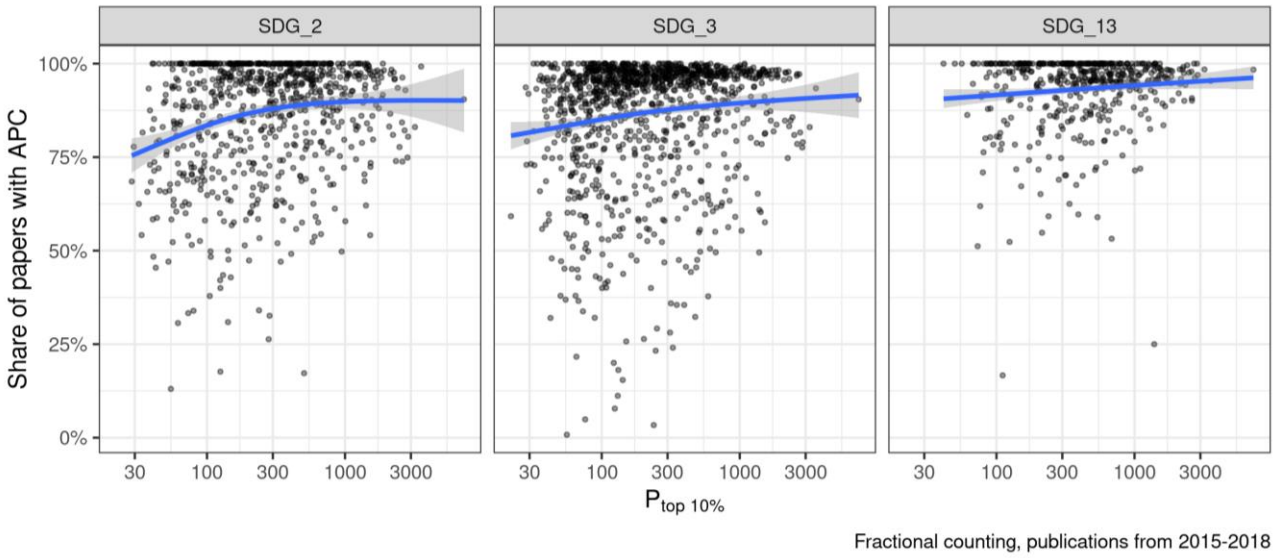

Figure 60: Association between institutional prestige and whether APCs are involved or not

In line with the above finding that the share of OA articles which involve an APC is stable, but with considerable variation across years, we find the same pattern when stratifying by institutional prestige. However, there are marked differences (Figure 61). Overall, higher ranking institutions publish more frequently with options involving APCs, while lower ranking institutions publish less frequently via the APC route. Given the significant variability visible, it is unclear whether the gap in terms of APC publishing between higher ranking and lower ranking institutions has been increasing or decreasing.

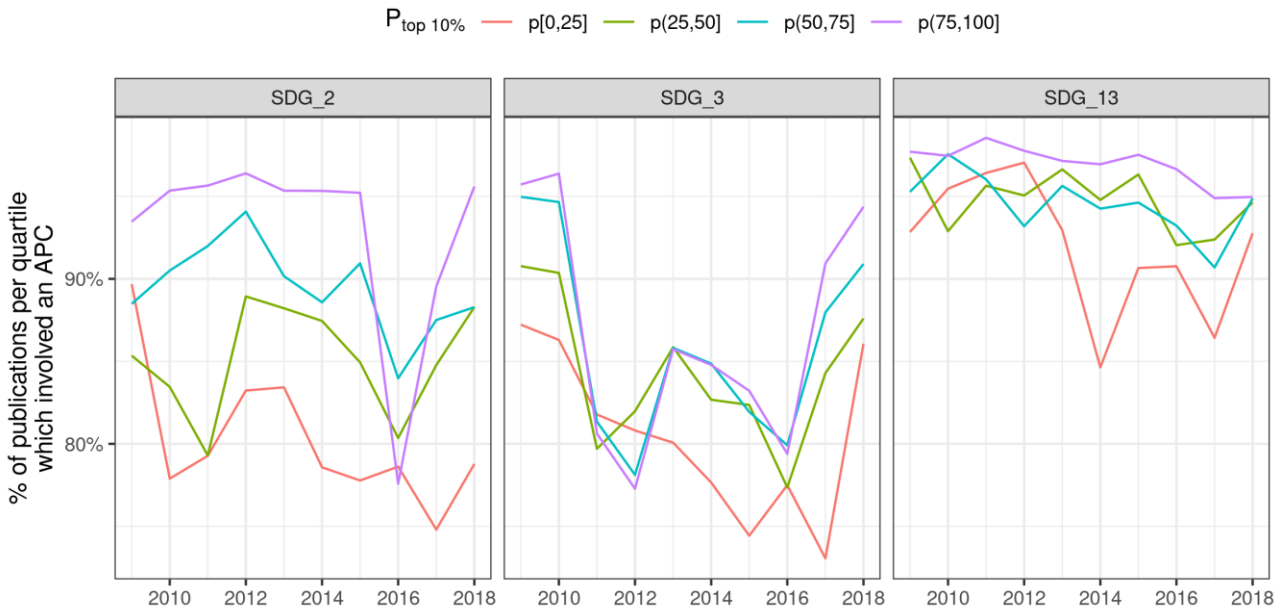

Figure 61: Shares of publications that involved an APC by institutional prestige

APC prices

We now move to the absolute values of APCs, still comparing them to institutional rankings. Here, we first include all journals that do not have an APC as having an APC of zero.[37] We find moderate positive correlations between institutional ranking and the average APC prices of the journals (Figure 62). In line with the above findings, SDG Zero Hunger (SDG 2) has the strongest association (first authors: .43; last authors: r = .46), SDG Climate Action (SDG 13) the lowest (first authors: r = .25; last authors: r = .22), with SDG Health/Well-Being (SDG 3) in the middle (first authors: r = .36; last authors: r = .35). Effects are very similar for first and last authors. We can therefore conclude that higher ranking institutions publish in journals involving higher APCs more frequently than lower ranking institutions.

[37] We repeated the analysis, excluding those journals that do not have APCs from the calculation of the means. Effects do not differ substantially from the current analysis (Figure A2 & Figure A3).

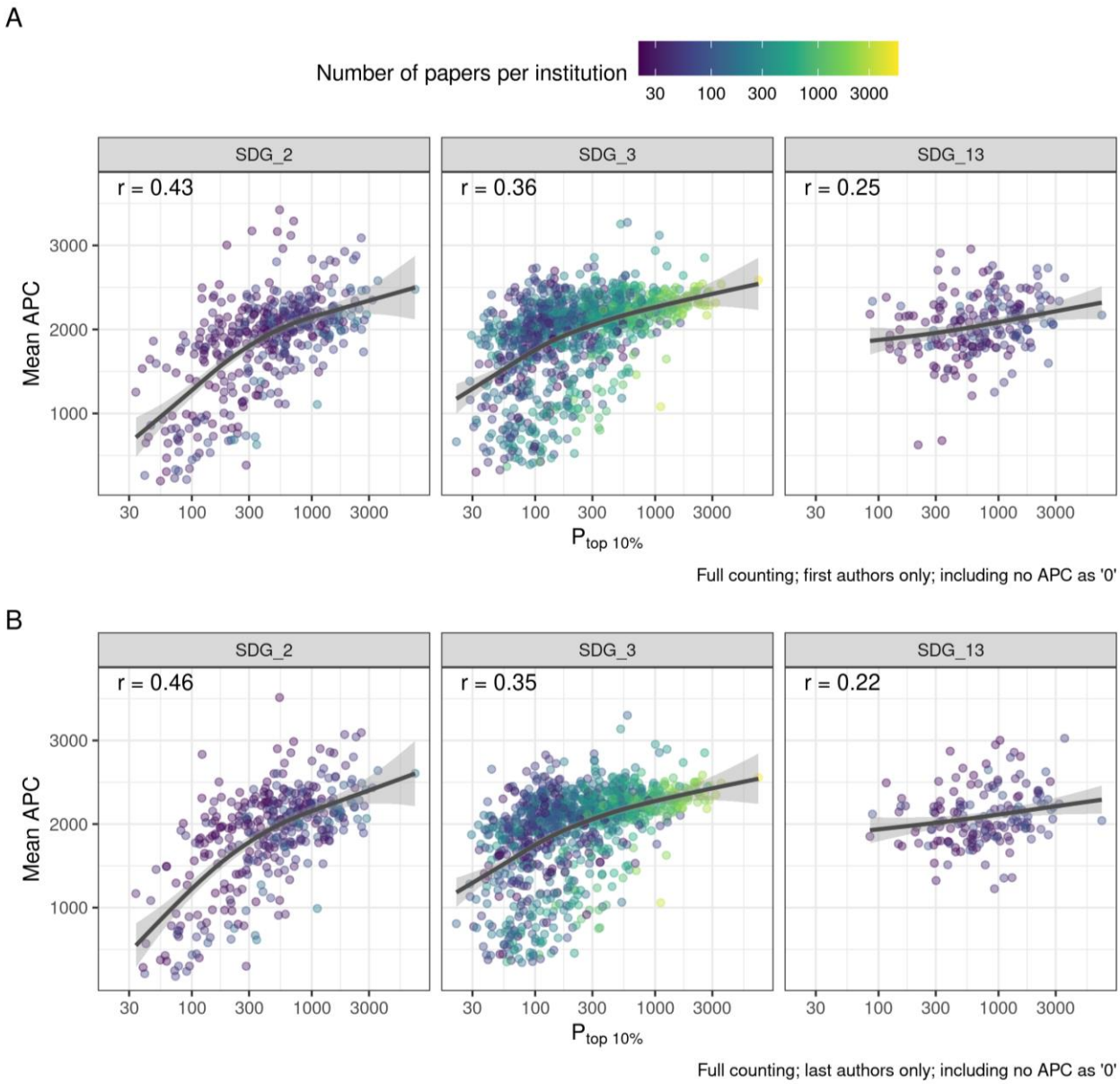

Figure 62: Mean APC per institution by institutional prestige (2015-2018). (A) First authors only. (B) Last authors only

To investigate how these associations develop over time, we analysed the median APC values of the journals for the quartiles of the ranking distribution (Figure 63). The lowest quartile exhibits no clear trend, but high variability over time, across all SDGs. All other quartiles exhibit an upward trend in the median APC prices of journals. This upward trend is strongest in the top quartile, across all SDGs.

Note that we do not have historical data on APC prices. The described effects therefore cannot be explained by rising APCs within journals, since we assume fixed APC prices per journal.[38] The most plausible explanation is that as subscription journals switch to the APC model and with new fully OA journals entering the publishing market, APC prices for prestigious journals increase (Gray 2020). Top ranking institutions continue to be able to cover APCs, thus maintaining their rate of APC funded OA, albeit with increasing prices. Lower ranking institutions, on the other hand, are unable to keep up with increasing APC prices, with their researchers

[38] OpenAPC (https://openapc.net/) has historical data on APCs, which could be leveraged in future analyses.

potentially being excluded from publishing in these high-APC journals. A further probable factor here is transformative agreements[39], which are likely more common in countries with a higher density of highly-ranking institutions.

Figure 63: Median journal APC fees for articles published by authors from all matched institutions in the Leiden ranking. (A) First authors only. (B) Last authors only

6.5. Discussion

The findings from this chapter paint a complex picture of the OA publishing landscape. Below, we discuss results related to a) general trends in OA publishing, b) OA publishing with respect to gender and academic age, c) OA publishing in terms of country, geographic region and world income group, and d) evolving patterns relating to the overall rate of OA publishing versus OA publishing prices (APCs).

Our findings on the rising share of OA publications over the last 15 years, with a peak share of OA publications of 52-58% depending on SDG are slightly higher than overall estimates found in previous studies (Piwowar et

[39] See section 6.2.3 on a brief description about what transformative agreements are.

al. 2018; Olejniczak and Wilson 2020). This is likely related to the disciplinary composition of our sample. In fact, Piwowar et al. (2018) found an OA share above 50% for biomedical research (2009-2015), which is in line with our findings of the highest share of OA publications among SDG Health/Well-Being (SDG 3).

In terms of hosting types of OA (journal only, repository only, or both) we found an increase in the share of articles that are hosted both by journals and repositories, and a marked decrease in the share of articles which are hosted solely by repositories. These findings cannot be directly compared to Piwowar et al. (2018) because of differing definitions. However, Piwowar et al. found a plateau effect for the share of green OA articles, which they explain by a) the shadowing of green OA by other pathways (gold, hybrid and bronze), because Unpaywall prefers publisher hosted over repository hosted, and b) the "backfilling" of green OA stocks since authors can make articles OA even years after publication of the article, and sometimes are mandated to do so only after a certain embargo period (Piwowar et al. 2018, 12). Since our methodology removes the influence of green OA that is shadowed by other pathways, the question remains whether the decrease in repository hosted OA can be fully explained by the slow backfilling of repository hosted content. Since the downward pattern is substantial (from above 30% in 2006 to below 20% in 2019) and stable over time, we would argue that this is not the case and we do in fact observe a decrease in repository-only hosted OA content.

Regarding academic age, we found lower rates of OA publishing among junior authors. This effect is consistent over time, across all three SDGs, and across author positions. The only difference is in terms of the effect's size, which is very small for first authors in SDG Health/Well-Being (SDG 3), and overall largest in SDG Climate Action (SDG 13), followed by SDG Zero Hunger (SDG 2), followed by SDG Health/Well-Being (SDG 3). Taking academic age as a proxy for seniority, these findings are consistent with Olejniczak and Wilson (2020) who also found higher rates of OA publishing among more senior academics. This can be interpreted in terms of OA publishing depending on resources which might not yet be available to younger researchers. Another potential explanation could be selection effects as described below for the case of gender might also play a role.

In terms of gender, we found the share of female authorships to be higher among OA articles than non-OA articles. These findings seem fairly remarkable, given that previous literature (Olejniczak and Wilson 2020) found that female authors tend to publish OA less often than men. It is important to note that these findings are compatible, and not in conflict with each other. Given the structure of our sample, we can only speak about the set of publications we sampled, and its properties. Since we did not sample all publications from the authors in our set, we cannot estimate the share of OA publications among all their publications, but just among SDG-related publications. A potential explanation for the higher share of female authorships among OA than non-OA publications would therefore be selection effects regarding the populations of OA and non-OA publications. For example, we found that higher-ranking institutions produce more OA, and also that higher-income countries produce more OA publications in absolute terms. If we assume that higher income countries and higher-ranking institutions (globally speaking) have a higher share of female authorships, this would explain the observed differences. Another dimension for this effect to occur would be along disciplines. In our analysis, we control for SDG area, but these areas cover a wide range of potential disciplines, which is why discipline could still be an explanatory factor in this regard. Finally, differences in methodologies (differing databases, gender definitions, etc.) could also be responsible for differences in results.

In terms of country differences, we found higher shares of OA publication among low income countries (LICs), particularly in SDG Health/Well-Being (SDG 3). For SDG Zero Hunger (SDG 2) and SDG Climate Action (SDG 13), the differences are smaller, with high income countries (HICs) exhibiting similar rates of OA publication to LICs. Lower medium income countries (LMICs) and upper medium income countries (UMICs) are lowest in SDG Zero Hunger and Climate Action. Our findings correspond to the results by Iyandemye and Thomas (2019), who found a higher share of OA publications among LICs in a sample of biomedical research publications. As discussed by Iyandemye and Thomas (2019), the higher share of OA publications among LICs might be explained by field-specific patterns of funding (targeted funding on HIV, Malaria, etc. in LICs). This hypothesis can be supported by our finding that LICs have similar or lower levels of OA publication than HICs in SDG Climate Action (SDG 13), which might not be subject to these specific funding streams.

Investigating differential rates of OA publishing according to institutional prestige within the analysed SDGs, we found a clear hierarchy: more prestigious institutions publish more OA than less prestigious institutions, which is in line with the findings from Siler et al. (2018). This effect is consistent across SDGs and also with the findings in Section 3.4.3, but not consistent over time. Across all three SDGs, but particularly in SDG Zero Hunger (SDG 2) and SDG Health/Well-Being (SDG 3), the association between institutional ranking and share of OA production is weakening. We can therefore conclude that albeit a first-mover advantage for better-resourced institutions to have higher rates of OA publishing is clearly visible, this advantage seems to diminish somewhat as OA publishing enters the mainstream.

The most salient findings of our study, however, relate to the stratification of OA publishing in terms of publishing costs (APCs). Overall, we found a high share of OA publishing (81.6% out of all articles where we had information from DOAJ, which was 45.6% of all publications in our sample) to include APCs. This share has remained at a similar level over the last 15 years, across all SDGs. Considering institutional prestige, we found that higher ranking institutions tend to publish more frequently in journals that charge an APC than lower ranking institutions. There is no clear indication whether this gap is closing or widening over time. At the same time, higher ranking institutions also publish more research in venues that charge higher APCs than lower ranking institutions. This difference has widened over time as well, with the least prestigious institutions (according to indicator $P_{top\ 10\%}$ from Leiden Ranking) publishing in journals with substantially lower APCs than the highest three quarters of the distribution.

These findings suggest that the APC publishing model has consequences for the question of who is able to contribute to the scientific record. Previous research (Gray 2020) has established a clear link between the APC publishing model (including APC prices) and common measures that are perceived as indicators for quality or prestige, like the impact factor, Eigenfactor, h-index or journal rank. Lower ranking institutions are thus increasingly precluded from publishing in outlets with the highest visibility and recognition. As noted above, inclusion into the Leiden ranking itself can already be understood as a marker for prestige, since a high production of internationally recognized research is a precondition. It is therefore fair to assume that the adverse effect of APC levels on the inclusivity and equality of the scholarly publishing landscape is even stronger for researchers from the least prestigious institutions.

7. Discussion

Open Science and RRI hold the promise to make scientific endeavours more inclusive, participatory, understandable, accessible, and re-usable for large audiences (Open Science) and more responsive to the needs and values of society (RRI). However, making processes open and responsible will not *per se* drive wide re-use or participation unless also accompanied by the capacity (in terms of knowledge, skills, financial resources, technological readiness and motivation) to do so. These capacities vary considerably across regions, institutions and demographics. Those advantaged by such factors will remain potentially privileged, putting Open Science and RRI's agendas of inclusivity at risk of propagating conditions of "cumulative advantage".

As Chin, Ribeiro, and Rairden (2019) remind us, "transitioning to open research involves significant financial costs." Open Science relies upon local training and support, as well as infrastructure and resources. Even in well-resourced regions such as Europe (Tenopir et al. 2017; MoRRI consortium 2018) and the US (Tenopir et al. 2014), readiness-levels of training and support infrastructure amongst nations and institutions are highly diverse. These disparities are, of course, even greater in what Siriwardhana (2015) terms "resource-poor" settings. Given that Open Science practices depend on underlying digital competences (Steinhardt 2020), the continuing realities of the digital divide (Maiti, Castellacci, and Melchior 2019) have real effects on participation in an Open Science world.

This work has focused on this crucial issue. It has sought to assess whether already prosperous countries and prestigious institutions are better placed and better resourced and hence more able to leverage the benefits of Open Science (especially Open Access) and RRI. It has done so by looking at the big picture across disciplines of production and consumption of OA globally and the impact of RRI policies in Europe. It has also sought to assess the specific impacts of OA publishing upon ON-MERRIT's three target domains of research relevant to the UN Sustainable Development Goals (SDGs): SDG 2 - Zero Hunger, SDG 3 - Good Health and Well-Being and SDG 13 - Climate Action.

7.1. Levels of resourcing and the uptake of RRI and Open Science

Throughout the analyses, in line with Ross-Hellauer et al. (2021), we find clear evidence that the more prestigious (and likely better resourced) institutions are better at adopting RRI principles and Open Science practices. They are capable of doing so more quickly and to a wider extent than their less well-resourced counterparts.

We find a strong correlation between the production of OA research papers and the consumption of OA research papers, and it is highly likely that this alignment is to some extent caused by accruing advantages due to early adoption of Open Science practices among better resourced institutions. Although identified more than 50 years ago, the effects of cumulative advantage are still perniciously present across academia. Furthermore, the size, wealth or location in academic networks of the institution in question might all influence how aspects of Open Science are taken up and to what extent benefits accrue.

However, what we found applicable for institutions does not necessarily hold for countries. In contrast to our initial assumption, we find no correlation between GDP per capita and OA consumption when considering all domains and countries. More specifically, we expected institutions situated in less developed countries to rely on OA resources in the papers they produce to a higher extent than more developed countries, but we did not find evidence of this. We tested this over three separate time periods between 2006 and 2020 and

found similar results for all periods. This is a somewhat surprising result indicating that the OA economy is much more strongly divided between prestigious and non-prestigious institutions than between the high-income and low-middle / low income countries.

These findings also stand in contrast to some of the goals attributed to the Open Science movement in fostering inclusivity and equity (Fecher and Friesike 2014). One key argument for Open Access to published research is that it can in principle be read by everyone. Randomized controlled trials of assigning OA or closed publishing have indeed found higher rates of views and downloads, compared to similar publications behind paywalls, but no citation advantage (Davis et al. 2008; Davis 2011). It might be plausible to assume that, regardless of any overall citation advantage, less well-resourced researchers rely more heavily on OA resources than more well-resourced actors. However, we find that among authors of research papers from lower income countries, there doesn't seem to be a preference for OA literature, indicating that these authors are likely able to gather (including possibly via illicit means like Sci-Hub) the research papers they need for their work. This does not diminish, of course, the possible impact of Open Access beyond academia. Enabling access to wider society (including the general public, industry, and policy-makers) remains a valuable contribution.

In line with our findings on the uptake of open scientific resources among policy-makers (c.f., ON-MERRIT Deliverables 5.2 and 5.3), these findings point to the fact that access alone is not sufficient for the uptake of (open) scientific resources, when it is not coupled with the necessary resources and skills. In other words, OA publishing does not seem to change how hierarchies in academia and scholarly communication operate. What is more, the change towards the APC model actually reinforces existing structural factors that drive cumulative advantage, as we examine next.

7.2. Stratification in publishing - APCs as a driving force

In Study 4 we demonstrate that higher ranking institutions publish more frequently in journals that charge an Article Processing Charge (APC) than lower ranking institutions. Higher ranking institutions also publish more research in venues that charge higher APCs than lower ranking institutions. Multiple factors might be contributing to these trends. Overall, there are three main sources for covering APCs: third-party funding, general institutional resources, and personal funds. It has been found that third party funding leads to higher rates of OA publishing (Larivière and Sugimoto 2018) and also to higher rates of APC-based OA (Olejniczak and Wilson 2020). Since it can be assumed that higher-ranking institutions generally also have higher rates of third-party funding, this is clearly a mechanism contributing to the observed effects. General institutional resources contribute to covering APCs in at least two ways: first through direct funding of APCs, and second through transformative agreements (Borrego, Anglada, and Abadal 2021), where institutions make deals with major publishers to cover APCs. It can be assumed that such deals are more common among higher ranking institutions with greater resourcing. Finally, researchers from industrialised nations rarely use personal funds to cover APCs, while this is quite common among researchers from developing nations (Björk and Solomon 2014). This likely reflects the diverging levels of institutional resourcing available for covering APCs.

These forces clearly perpetuate the system of cumulative advantage (CA) inherent to academia, as well-funded research groups are better able to secure OA publications in prestigious journals with high APCs, leading to citation advantages and further funding down the line. We believe that this demonstrates the impact of APC pricing on the scholarly landscape and that these charges may have a chilling effect on opportunity and equality for researchers from less prestigious or less wealthy institutions. Such stratifications in publishing, favouring traditionally-advantaged actors, will only exacerbate historical inequalities (Garuba

2013) and undermine wider aims of Open Science. Hence, as Nyamnjoh argues, for "open access to be meaningful … questions of content and the epistemological, conceptual, methodological and contextual specificities that determine or impinge upon it are crucial" (Nyamnjoh 2010). We therefore agree with Czerniewicz (2015) who argues that such consequences are the result of too narrow a focus on achieving OA per se, by whichever means, without acknowledging "the inequitable global power dynamics of global knowledge production and exchange". Rather, she suggests, we must broaden our focus "from access to knowledge to full participation in knowledge creation and in scholarly communication".

7.3. Individual-level demographics

In addition to considering the overall picture at the institutional and national levels, our investigation also examined two characteristics specific to individual researchers, namely academic age (defined as length of time since first publication) and gender.

First, in regards to academic age, we found lower rates of OA publishing among junior authors. This effect is consistent over time, across disciplines, and across author positions. While further investigation needs to be undertaken to establish potential reasons for these dynamics, initial hypotheses could be formed based on the premise of availability of resources. Junior researchers might have lower rates of third-party funding, which has been found to be associated with higher rates of OA publishing (Olejniczak and Wilson 2020). It may also be the case that junior researchers, whose conditions of employment are among the most precarious in academia, are simply less inclined to take up OA, preferring to publish in the traditional "high impact" journals whose name-brand value still unfortunately bears so strongly upon outcomes in promotion, review and tenure processes (c.f., ON-MERRIT D6.1 "Investigating Institutional Structures of Reward and Recognition for Open Science & RRI").

Second, this study investigated the role of gender across a broad spectrum; from the policy level, using data from the SuperMoRRI project to the publication level, using data from MAG. We studied the impact of gender using the share of female authorships on research articles published in the three domains related to the UN SDGs. We find an encouraging upwards trend (although gender parity remains still very distant), with the share increasing from 19-30% in 2006 to 27-37% in 2019, the exact percentage being domain dependent. We found the share of female authorships to be the highest in SDG Health/Well-Being (SDG 3), and the lowest in SDG Climate Action (SDG 13).

In Study 2, we were surprised to find a lack of correlation between how a country performs in terms of gender equality, measured using the RRI indicators (*F_gender*), and the actual balance in the numbers of male / female researchers (*F_F_MF*) in that country. In terms of the embedding of gender specific policies and practices, there remain large differences across continental Europe. In particular, we find that the new EU13 countries perform far less well than their western counterparts. A range of sociological factors might influence this apparent gap between policies and structural composition of academic labour. First, others have noted that higher shares of female researchers among institutions from Eastern Europe and South America might be due to lower wages for "science jobs" in these regions, leading men to pursue careers in other economic sectors or in other countries (Guglielmi 2019). Furthermore, Larivière et al. (2013) note that greater gender parity in Eastern European countries might be due to their history of communism. This might be due to higher educational attainment among women, compared to other countries, or due to further structural or cultural factors. The relationship between gender policies and outcomes in terms of gender parity is therefore complex and not easily distilled into recommendations for further policies.

One encouraging result shows a strong overall correlation between measures of *public engagement* with science (*F_engagement*), and RRI policies at the national level (r=0.79, n=344). We see demonstrably higher levels of public engagement with science in countries where these policies are more embedded. However, it is unclear to what extent public engagement with science is the result of, or the precondition for, implementing RRI policies.

7.4. Open Science, RRI and the SDGs

As argued by Albornoz et al. (2020), Open Science policies are situated within power imbalances and historical inequalities with respect to knowledge production (c.f. Mirowski 2018). Uncritical narratives of openness therefore may fail to address structural barriers in knowledge production and hence perpetuate the cumulative advantage of dominant groups and the knowledge they produced.

While our global macro analysis in Study 1 does not show a strong correlation between OA Production rates and GDP per capita, it is worthwhile noting that, as revealed by Study 4, the situation might be different within specific disciplines. Analysing the uptake of OA publishing across countries and SDGs, we find higher rates of OA publishing amongst low income and high income countries than in medium income countries. This is especially true for SDG Health/Well-Being (SDG 3), but less so for SDG Climate Action (SDG 13).

The SDGs are global problems requiring global consensus and global solutions. Concentration of research into these key areas amongst particular well-resourced actors is dangerous. As discussed in ON-MERRIT D5.3, the question of who is performing research, under which cultural presuppositions or towards which political ends, is ever-pressing. If research on the SDGs is dominated by traditional actors, especially those in the Global North, then it is possible that priorities are set based on particularistic perspectives. As research is increasingly mission-driven, with national funding directed at national policy goals or national interpretations of global goals like the SDGs, diversity and equity of participation in the global effort to address the SDGs is threatened. Effects of cumulative advantage in terms of who contributes will hence potentially lead to the priorities of those home-countries dominating the conversation.

7.5. Implications

There is a potential response to these concerns that questions whether it really is problematic that actors with more resources are quicker to take up new practices, since they are also the ones who bear the cost of learning the lessons such that less well-resourced actors can later more easily implement them and catch up.

To this, we would answer that the danger is two-fold: (1) As OS and RRI become the norm, actors with lower resources and slower uptake will, in line with the principles of cumulative advantage, be further penalised for not conforming to current OS/RRI standards. The environment will already have changed such that open and responsible practices are expected before poorer institutions have the capacity to properly engage with them (thus risking further exclusion). As the more prestigious institutions are better able to attract talent, funding, etc., the process will reinforce itself; (2) If well-resourced actors are quickest to act, they may (even unknowingly) shape the OS/RRI environment in ways which suit their own interests and levels of resourcing (e.g., APC-funded OA). This can create systemic discrimination for those who are not moving forward now. Hence, we would argue that not only have we shown that a problem exists, but that answering it is pressing, particularly for those currently left behind

Across all of these studies we have investigated the ways in which prevailing capacities, resources and network centralities – combined with structural inequalities and biases – can shape Open Science and RRI outcomes. Inequalities and dynamics of cumulative advantage pervade modern scholarship, and our results show that despite its potential to improve equity in many areas, Open Science and RRI are not exempt. Cumulative advantage relates to logics of accumulation and preferential attachment based on network positionality and possession of resources. The resource-intensive and networked nature of Open Science and RRI mean they are also vulnerable to these logics. Explicitly linking authorship channels to possession of resources potentially stratifies Open Access publishing. The expensive infrastructures and training necessary to participate in engaging with OS/RRI practices means those privileged in these regards are primed to benefit most, at least initially. The importance of such underlying competencies means that ensuring access is not enough to ensure equity of opportunity in an Open Science world absent of broader measures to overcome the digital divide. Merton advises that cumulative advantage directs our attention to "the ways in which initial comparative advantages of trained capacity, structural location, and available resources make for successive increments of advantage such that the gaps between the haves and the have-nots in science (as in other domains of social life) widen until dampened by countervailing processes" (Merton 1988). Having identified ways in which cumulative advantage may be at play, it is our next task to identify the countervailing processes which might mitigate their ill effects. In our final section, we present a summary of our overall conclusions and make provisional recommendations (to be refined in later ON-MERRIT activities) for such corrective measures.

8. Conclusions

Scientific knowledge is a key resource for achieving societal and economic goals. Open Science and RRI promise to fundamentally transform scholarship to bring greater transparency, inclusivity and participation to research processes, and increase the academic, economic and societal impact of research outputs. These form a cross-cutting agenda that stands to contribute to most of the UN's Sustainable Development Goals, as well as being a central pillar of the European Commission's Digital Single Market strategy. Yet access to scientific products and processes is not made uniform simply because they are made available via the Internet. How equitable is implementation of OS and RRI across a range of stakeholder categories, and in particular for those at the peripheries? Might RRI interventions in some cases actually deepen socio-economic inequalities (the digital divide) and be at conflict with wider sustainable development goals? How do geographical, socio-economic, cultural and structural conditions lead to peripheral configurations in the European knowledge landscape? What factors are at play and what can be done (at a policy level) to foster absorptive capacity and enhance OS/RRI uptake and contributions to scientific production across regions?

Such questions lie at the heart of ON-MERRIT. This work constitutes a major part of our attempt to answer them. In order to investigate these effects of cumulative advantage in the transition to Open Science and RRI, we took both broad and focused approaches to conduct four complementary studies into Open Access publishing, the uptake of RRI policies, and their relationships to the UN Sustainable Development Goals.

In Study 1 (section 3 of this document), we investigated by whom Open Access publications are produced, and who cites them, on a global scale. Contrary to initial hypotheses, we find a strong correlation between the shares of OA publication and OA citation among institutions, but not across countries. Combining data on the uptake of RRI policies with bibliometric information, Study 2 (section 4) finds low uptake of RRI policies and practices in the new EU countries (EU 13), a strong correlation between RRI policies at the national level and measures of public engagement with science, as well as no correlation between gender equality policies and the actual balance of male vs female researchers.

Investigating research published on three key UN Sustainable Development Goals (SDG Zero Hunger, SDG Health/Well-Being, SDG Climate Action), Study 3 (section 5) finds persistent institutional stratification, a rising share of female authorship and an overall increase in the research being published on the three SDGs. Building on Study 3, in Study 4 (section 6) we investigate the uptake of OA publishing among the three key SDGs in terms of institutions, countries, and individual demographics. We find that well-resourced actors publish more frequently OA in the SDG areas, as well as publishing in journals with on average higher APCs.

The four studies presented in this deliverable combine to highlight that it is the higher ranked, more prosperous and more prestigious institutions that appear best able to adopt, adapt to, and benefit from, the evolving landscape of Open Access publishing. These trends hold true over time, on the global level, and when broken down to individual continents and subject areas (SDGs). Persistent structural inequalities in contemporary academic publishing are not necessarily remedied by the Open Science movement, with specific trends such as APC-driven OA publishing potentially exacerbating dynamics of cumulative advantage. If research on key global issues is only driven by well-resourced actors, it risks being oblivious to challenges faced by societies and communities less embedded into the global production of knowledge.

8.1. Limitations

The results presented in this deliverable are subject to multiple limitations. *First*, our analyses are bound by their underlying data. We used diverse data sources, with differing contexts of origin and varying conceptualisations. There might be some ambiguity in how we merged datasets such as MAG and the Leiden Ranking, with slightly differing coverages and diverging conceptualisations of which entities belong to a certain university. *Second*, the selection of publications which are relevant for tackling the three key SDGs of our study (Zero Hunger (SDG 2), Health/Well-Being (SDG 3), and Climate Action (SDG 13) is based on an approach that is still in development. We are confident that the results presented in this deliverable are meaningful, but better algorithms could provide a more precise sample of SDG-relevant research and thus more valid conclusions. *Third*, both the analysis of OA production and consumption, as well as the analysis of OA and APC publishing across SDGs are only a first step. Further analyses need to be undertaken to model outcomes more directly, incorporating multiple factors into a single explanatory model.

8.2. Recommendations

Building on the discussion presented above, we offer the following set of preliminary recommendations to OS/RRI policy-makers (funders and governments), academic and scientific institutional leaders, and researchers. (These recommendations will be refined during the synthesis phase of ON-MERRIT, which will include co-creation processes with these three stakeholder groups.)

Science policymakers (e.g., funders/governments) should:

- Support less prestigious institutions in building OS and RRI capacity and awareness. Our results show that those who produce more OA tend also to benefit more, hence it is important to create a level playing-field and close the significant difference between institutions that we observed.
- We recommend waiving APCs, not just to less-developed countries, but also to less prestigious institutions in more developed countries. This follows our observation that the less prestigious institutions are those that are on average slower in investing into and therefore also reaping the benefits of OA.
- Commit resources to support, in a sustainable manner, key OS infrastructure initiatives delivering freely available resources that provide the foundations for conducting science on science. For instance, the recent announcement that Microsoft will stop offering Microsoft Academic Graph (MAG) from January 2022, i.e. one of the key datasets on which this study is based, is concerning. We need open alternatives for these infrastructures.
- Support the creation of responsible metrics for OS practices that can be obtained at the granularity of individual researchers and that are related to the rigour of their scholarship. While open peer review, reproducibility, data citations, considering the meaning and motivation of citations, etc. all provide promising avenues in this direction, assessing and understanding how these are and should be linked to academic success is still in its infancy.
- Mandate that research performing organisations annually release RRI data along a required set of dimensions. Doing so should be a precondition for participating in research grants. As an example, this is practiced in the UK in relation to the ATHENA SWAN accreditation scheme, which promotes good practices in higher education and research institutions towards the advancement of gender equality.

- Closely monitor who is contributing to mission-driven research such as that on the SDGs, and assessment of the ways in which global perspectives may be excluded through the dominance of specific actors.

Research performing organisations should:

- Regularly collect and make publicly available RRI data, ideally at the level of faculties/departments. The fact that we were only able to obtain RRI data at a country level and that they are not regularly collected, seriously limits the ability of monitoring progress towards RRI and reduces the potential of research into its benefits.
- Publicly and annually release information about how much they spend for APCs and subscriptions to academic literature.
- Dedicate, via their academic libraries, a set proportion of their budgets for the support of a range of open, shared and not-for-profit scholarly infrastructures. These could include but are not limited to funding of non-APC consortial "diamond" models of OA where neither author or reader must pay directly, promoting library publishing to combat APC stratification, supporting open citation, and new, emerging and existing scholarly scholarly data infrastructures.

Researchers should:

- Deposit their Author Accepted Manuscripts (AAMs) into repositories. This is especially important in situations where the researcher, or their institution, cannot afford to pay an APC. A strong and somewhat selfish incentive to researchers should be that by making their research OA, they are primarily helping themselves in gaining cumulative advantage over their peers who don't. As an (important) side-effect, they also benefit other researchers, professionals and the general public.
- Try to actively avoid biases, including being aware of the risk of unconscious biases, in decisions including but not limited to the selection of publication venue and institutional prestige or location of collaborating partners. Researchers should aim to take these decisions on merit, to limit the effects of cumulative advantage we studied.

9. References

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*.

Albornoz, D., M. Huang, I.M. Martin, M. Mateus, A.Y. Touré, and L. Chan. 2020. "Framing Power: Tracing Key Discourses in Open Science Policies." In *22nd International Conference on Electronic Publishing - Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure, ELPUB 2018*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085471179&partnerID=40&md5=4d9a26ae9080a4c38e819ac062070c9c.

Allison, Paul D., and J. Scott Long. 1990. "Departmental Effects on Scientific Productivity." *American Sociological Review* 55 (4): 469–78. https://doi.org/10.2307/2095801.

Armitage, Caroline S., Marta Lorenz, and Susanne Mikki. 2020. "Mapping Scholarly Publications Related to the Sustainable Development Goals: Do Independent Bibliometric Approaches Get the Same Results?" *Quantitative Science Studies* 1 (3): 1092–1108. https://doi.org/10.1162/qss_a_00071.

Athena Research & Innovation Center, Directorate-General for Research and Innovation (European Commission), PPMI, and UNU-MERIT. 2021. *Monitoring the Open Access Policy of Horizon 2020: Final Report*. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2777/268348.

Ball, Philip. 2002. "Paper Trail Reveals References Go Unread by Citing Authors." *Nature* 420 (6916): 594–594. https://doi.org/10.1038/420594a.

Bautista-Puig, Núria, Ana Marta Aleixo, Susana Leal, Ulisses Azeiteiro, and Rodrigo Costas. 2021. "Unveiling the Research Landscape of Sustainable Development Goals and Their Inclusion in Higher Education Institutions and Research Centers: Major Trends in 2000–2017." *Frontiers in Sustainability* 2: 12. https://doi.org/10.3389/frsus.2021.620743.

Björk, Bo-Christer, and David Solomon. 2014. "Developing an Effective Market for Open Access Article Processing Charges." Zenodo. https://doi.org/10.5281/zenodo.51788.

Bohannon, John. 2016. "Who's Downloading Pirated Papers? Everyone." *Science* 352 (6285): 508–12. https://doi.org/10.1126/science.352.6285.508.

Bornmann, Lutz, and Hans-Dieter Daniel. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behavior." *Journal of Documentation* 64 (1): 45–80. https://doi.org/10.1108/00220410810844150.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel. 2013. "Multilevel-Statistical Reformulation of Citation-Based University Rankings: The Leiden Ranking 2011/2012." *Journal of the American Society for Information Science and Technology* 64 (8): 1649–58.

Borrego, Ángel, Lluís Anglada, and Ernest Abadal. 2021. "Transformative Agreements: Do They Pave the Way to Open Access?" *Learned Publishing* 34 (2): 216–32. https://doi.org/10.1002/leap.1347.

Brankovic, Jelena. 2021. "Why Rankings Appear Natural (But Aren't)." *Business & Society*, May, 000765032110156. https://doi.org/10.1177/00076503211015638.

Burchardt, Jørgen. 2014. "Researchers Outside APC-Financed Open Access: Implications for Scholars Without a Paying Institution." *SAGE Open* 4 (4): 2158244014551714. https://doi.org/10.1177/2158244014551714.

Chin, Jason M., Gianni Ribeiro, and Alicia Rairden. 2019. "Open Forensic Science." *Journal of Law and the Biosciences* 6 (1): 255–88. https://doi.org/10.1093/jlb/lsz009.

Cole, Jonathan R., and Stephen Cole. 1973. *Social Stratification in Science*. Chicago, Ill.: University of Chicago Press.

Cole, Stephen. 1979. "Age and Scientific Performance." *American Journal of Sociology* 84 (4): 958–77. https://doi.org/10.1086/226868.

Collyer, Fran M. 2018. "Global Patterns in the Publishing of Academic Knowledge: Global North, Global South." *Current Sociology* 66 (1): 56–73. https://doi.org/10.1177/0011392116680020.

Costas, Rodrigo, and María Bordons. 2011. "Do Age and Professional Rank Influence the Order of Authorship in Scientific Publications? Some Evidence from a Micro-Level Perspective." *Scientometrics* 88 (1): 145–61. https://doi.org/10.1007/s11192-011-0368-z.

Costas, Rodrigo, Thed N. van Leeuwen, and María Bordons. 2010. "A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level: The Effects of Age on Productivity and Impact." *Journal of the American Society for Information Science and Technology* 61 (8): 1564–81. https://doi.org/10.1002/asi.21348.

Czerniewicz, L. 2015. "Opinion: Confronting Inequitable Power Dynamics of Global Knowledge Production and Exchange." *Water Wheel* 14 (5): 26–29.

Davis, Philip M. 2011. "Open Access, Readership, Citations: A Randomized Controlled Trial of Scientific Journal Publishing." *The FASEB Journal* 25 (7): 2129–34. https://doi.org/10.1096/fj.11-183988.

Davis, Philip M., Bruce V. Lewenstein, Daniel H. Simon, James G. Booth, and Mathew J. L. Connolly. 2008. "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial." *BMJ* 337 (July): a568. https://doi.org/10.1136/bmj.a568.

Dehon, Catherine, Alice McCathie, and Vincenzo Verardi. 2010. "Uncovering Excellence in Academic Rankings: A Closer Look at the Shanghai Ranking." *Scientometrics* 83 (2): 515–24.

Demeter, M., and R. Istratii. 2020. "Scrutinising What Open Access Journals Mean for Global Inequalities." *Publishing Research Quarterly*. https://doi.org/10.1007/s12109-020-09771-9.

Dion, Michelle L., Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. "Gendered Citation Patterns across Political Science and Social Science Methodology Fields." *Political Analysis* 26 (3): 312–27. https://doi.org/10.1017/pan.2018.12.

Donner, Paul, Christine Rimmert, and Nees Jan van Eck. 2020. "Comparing Institutional-Level Bibliometric Research Performance Indicator Values Based on Different Affiliation Disambiguation Systems." *Quantitative Science Studies* 1 (1): 150–70. https://doi.org/10.1162/qss_a_00013.

Dorta-González, Pablo, Sara M. González-Betancor, and María Isabel Dorta-González. 2017. "Reconsidering the Gold Open Access Citation Advantage Postulate in a Multidisciplinary Context: An Analysis of the Subject Categories in the Web of Science Database 2009–2014." *Scientometrics* 112 (2): 877–901. https://doi.org/10.1007/s11192-017-2422-y.

Drori, Gili S., John W. Meyer, Francisco O. Ramirez, and Evan Schofer. 2003. *Science in the Modern World Polity: Institutionalization and Globalization*. Stanford (Calif.): Stanford University Press.

Ellers, J., T.W. Crowther, and J.A. Harvey. 2017. "Gold Open Access Publishing in Mega-Journals: Developing Countries Pay the Price of Western Premium Academic Output." *Journal of Scholarly Publishing* 49 (1): 89–102. https://doi.org/10.3138/jsp.49.1.89.

ElSabry, ElHassan. 2017. "Unaffiliated Researchers: A Preliminary Study." *Challenges* 8 (2): 20. https://doi.org/10.3390/challe8020020.

European Commission and Directorate-General for Research and Innovation. 2012. *Responsible Research and Innovation: Europe's Ability to Respond to Societal Challenges.*

Fecher, Benedikt, and Sascha Friesike. 2014. "Open Science: One Term, Five Schools of Thought." In *Opening Science*, edited by Sönke Bartling and Sascha Friesike, 17–47. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, et al. 2018. "Science of Science." *Science* 359 (6379). https://doi.org/10.1126/science.aao0185.

Fox, Mary Frank. 1983. "Publication Productivity among Scientists: A Critical Review." *Social Studies of Science* 13 (2): 285–305. https://doi.org/10.1177/030631283013002005.

Frenken, Koen, Gaston J. Heimeriks, and Jarno Hoekman. 2017. "What Drives University Research Performance? An Analysis Using the CWTS Leiden Ranking Data." *Journal of Informetrics* 11 (3): 859–72. https://doi.org/10.1016/j.joi.2017.06.006.

Gadd, Elizabeth. 2021. "Mis-Measuring Our Universities: Why Global University Rankings Don't Add Up." *Frontiers in Research Metrics and Analytics* 6 (September): 680023. https://doi.org/10.3389/frma.2021.680023.

Garuba, A.R. 2013. "The Prospects of Bridging the Digital Divide in Africa." *Library Philosophy and Practice* 2013. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085323161&partnerID=40&md5=530601d0b82a6c389c300e61b548f11a.

Gonzalez Garcia, Erika, Ernesto Colomo Magana, and Andrea Civico Ariza. 2020. "Quality Education as a

Sustainable Development Goal in the Context of 2030 Agenda: Bibliometric Approach." *Sustainability* 12 (15): 5884. https://doi.org/10.3390/su12155884.

Gray, R.J. 2020. "Sorry, We're Open: Golden Open-Access and Inequality in Non-Human Biological Sciences." *Scientometrics* 124 (2): 1663–75. https://doi.org/10.1007/s11192-020-03540-3.

Guglielmi, Giorgia. 2019. "Eastern European Universities Score Highly in University Gender Ranking." *Nature*, May. https://doi.org/10.1038/d41586-019-01642-4.

Hajjem, Chawki, S. Harnad, and Y. Gingras. 2005. "Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How It Increases Research Citation Impact." *IEEE Data Eng. Bull.*

Harnad, S., and Tim Brody. 2004. "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals." *D-Lib MAgazine*. https://www.semanticscholar.org/paper/Comparing-the-Impact-of-Open-Access-(OA)-vs.-Non-OA-Harnad-Brody/7552ed5a3960106909a51951108fded284c1aeed.

Hart, Kamber L., Sophia Frangou, and Roy H. Perlis. 2019. "Gender Trends in Authorship in Psychiatry Journals From 2008 to 2018." *Biological Psychiatry* 86 (8): 639–46. https://doi.org/10.1016/j.biopsych.2019.02.010.

Harzing, Anne-Wil, and Satu Alakangas. 2017. "Microsoft Academic: Is the Phoenix Getting Wings?" *Scientometrics* 110 (1): 371–83.

Helgesson, Gert, and Stefan Eriksson. 2019. "Authorship Order." *Learned Publishing* 32 (2): 106–12. https://doi.org/10.1002/leap.1191.

Holmberg, Kim, Juha Hedman, Timothy D. Bowman, Fereshteh Didegah, and Mikael Laakso. 2020. "Do Articles in Open Access Journals Have More Frequent Altmetric Activity than Articles in Subscription-Based Journals? An Investigation of the Research Output of Finnish Universities." *Scientometrics* 122 (1): 645–59. https://doi.org/10.1007/s11192-019-03301-x.

Huang, Chun-Kai (Karl), Cameron Neylon, Richard Hosking, Lucy Montgomery, Katie S Wilson, Alkim Ozaygen, and Chloe Brookes-Kenworthy. 2020. "Evaluating the Impact of Open Access Policies on Research Institutions." Edited by Julia Deathridge, Peter Rodgers, and Bianca Kramer. *ELife* 9 (September): e57067. https://doi.org/10.7554/eLife.57067.

Hug, Sven E., Michael Ochsner, and Martin P. Brändle. 2017. "Citation Analysis with Microsoft Academic." *Scientometrics* 111 (1): 371–78. https://doi.org/10.1007/s11192-017-2247-8.

Iefremova, Olesia, Kamil Wais, and Marcin Kozak. 2018. "Biographical Articles in Scientific Literature: Analysis of Articles Indexed in Web of Science." *Scientometrics* 117 (3): 1695–1719. https://doi.org/10.1007/s11192-018-2923-3.

Iyandemye, Jonathan, and Marshall P. Thomas. 2019. "Low Income Countries Have the Highest Percentages of Open Access Publication: A Systematic Computational Analysis of the Biomedical Literature." *PLOS ONE* 14 (7): e0220229. https://doi.org/10.1371/journal.pone.0220229.

Jayabalasingham, Bamini, Roy Boverhof, Kevin Agnew, and L. Klein. 2019. "Identifying Research Supporting the United Nations Sustainable Development Goals" 1 (October). https://doi.org/10.17632/87txkw7khs.1.

King, Molly M., Carl T. Bergstrom, Shelley J. Correll, Jennifer Jacquet, and Jevin D. West. 2017. "Men Set Their Own Cites High: Gender and Self-Citation across Fields and over Time." *Socius* 3 (January): 2378023117738903. https://doi.org/10.1177/2378023117738903.

Kousha, Kayvan, and Mahshid Abdoli. 2010. "The Citation Impact of Open Access Agricultural Research: A Comparison between OA and Non-OA Publications." *Online Information Review* 34 (5): 772–85. https://doi.org/10.1108/14684521011084618.

Kuhn, Thomas S. 2012. *The Structure of Scientific Revolutions*. Fourth edition. Chicago ; London: The University of Chicago Press.

Kuntal, Bhusan K, Tarini Shankar Ghosh, and Sharmila S Mande. 2014. "Igloo-Plot: A Tool for Visualization of Multidimensional Datasets." *Genomics* 103 (1): 11–20.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. "Bibliometrics: Global Gender Disparities in Science." *Nature* 504 (7479): 211–13. https://doi.org/10.1038/504211a.

Larivière, Vincent, and Cassidy R. Sugimoto. 2018. "Do Authors Comply When Funders Enforce Open Access to Research?" *Nature* 562 (7728): 483–86. https://doi.org/10.1038/d41586-018-07101-w.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts: With a New Postscript and Index by the Authors*. 1. Princeton paperback printing. Princeton, N.J: Princeton University Press.

Leeuwen, Thed van. 2019. "The Matthew Effect of Plan S: Is Gold OA Publishing Mainly a Business Model Fitting the Rich in Science?" Presented at the Pubmet 2019, Zadar, Croatia. https://meduza.carnet.hr/index.php/media/watch/13767.

Leeuwen, Thed van, Clifford Tatum, and Paul Wouters. 2015. "Open Access Publishing and Citation Impact - An International Study." In *Proceedings of the 15th International Conference on Scientometrics and Informetrics, Istanbul, Turkey, June 29 - July 3, 2015*, edited by Albert Ali Salah, Yasar Tonta, Alkim Almila Akdag Salah, Cassidy R. Sugimoto, and Umut Al. ISSI Society.

Maiti, D., F. Castellacci, and A. Melchior. 2019. *Digitalisation and Development: Issues for India and Beyond*. *Digitalisation and Development: Issues for India and Beyond*. https://doi.org/10.1007/978-981-13-9996-1.

Mallapaty, Smriti. 2020. "China Bans Cash Rewards for Publishing Papers." *Nature* 579 (7797): 18–18. https://doi.org/10.1038/d41586-020-00574-8.

Matheka, Duncan Mwangangi, Joseph Nderitu, Daniel Mutonga, Mary Iwaret Otiti, Karen Siegel, and Alessandro Rhyll Demaio. 2014. "Open Access: Academic Publishing and Its Implications for Knowledge Equity in Kenya." *Globalization and Health* 10 (1): 26. https://doi.org/10.1186/1744-8603-10-26.

Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." *Science* 159 (3810): 56–63. https://doi.org/10.1126/science.159.3810.56.

———. 1988. "The Matthew Effect in Science, II. Cumulative Advantage and the Symbolism of Intellectual Property." *Isis* 79: 606–23.

Meschede, Christine. 2020. "The Sustainable Development Goals in Scientific Literature: A Bibliometric Overview at the Meta-Level." *Sustainability* 12 (11): 4461. https://doi.org/10.3390/su12114461.

Mirowski, Philip. 2018. "The Future(s) of Open Science." *Social Studies of Science* 48 (2): 171–203. https://doi.org/10.1177/0306312718772086.

Moed, Henk F. 2017. "A Critical Comparative Analysis of Five World University Rankings." *Scientometrics* 110 (2): 967–90.

MoRRI consortium. 2018. "Final Report – Summarising Insights from the MoRRI Project." http://morri-project.eu/reports/2018-05-24-final-report-summarising-insights-from-the-morri-project.

Nazari, Mateus Torres, Janaina Mazutti, Luana Girardi Basso, Luciane Maria Colla, and Luciana Brandli. 2020. "Biofuels and Their Connections with the Sustainable Development Goals: A Bibliometric and Systematic Review." *Environment Development and Sustainability*. https://doi.org/10.1007/s10668-020-01110-4.

Nicholas, David, Chérifa Boukacem-Zeghmouri, Jie Xu, Eti Herman, David Clark, Abdullah Abrizah, Blanca Rodríguez-Bravo, and Marzena Świgoń. 2019. "Sci-Hub: The New and Ultimate Disruptor? View from the Front." *Learned Publishing* 32 (2): 147–53. https://doi.org/10.1002/leap.1206.

Nyamnjoh, Francis. 2010. "Institutional Review: Open Access and Open Knowledge Production Processes: Lessons from CODESRIA," February. https://doi.org/10.23962/10539/19772.

Olejniczak, Anthony J., and Molly J. Wilson. 2020. "Who's Writing Open Access (OA) Articles? Characteristics of OA Authors at Ph.D.-Granting Institutions in the United States." *Quantitative Science Studies*, October, 1–22. https://doi.org/10.1162/qss_a_00091.

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. 2018. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles." *PeerJ* 6 (February): e4375. https://doi.org/10.7717/peerj.4375.

Pizzi, Simone, Andrea Caputo, Antonio Corvino, and Andrea Venturelli. 2020. "Management Research and the UN Sustainable Development Goals (SDGs): A Bibliometric Investigation and Systematic Review." *Journal of Cleaner Production* 276 (December): 124033. https://doi.org/10.1016/j.jclepro.2020.124033.

Pukelis, Lukas, Nuria Bautista Puig, Mykola Skrynik, and Vilius Stanciauskas. 2020. "OSDG -- Open-Source

Approach to Classify Text Data by UN Sustainable Development Goals (SDGs).” *ArXiv:2005.14569 [Cs]*, May. http://arxiv.org/abs/2005.14569.

Pusser, Brian, and Simon Marginson. 2013. “University Rankings in Critical Perspective.” *The Journal of Higher Education* 84 (4): 544–68.

Ramírez-Castañeda, Valeria. 2020. “Disadvantages in Preparing and Publishing Scientific Papers Caused by the Dominance of the English Language in Science: The Case of Colombian Researchers in Biological Sciences.” *PLOS ONE* 15 (9): e0238372. https://doi.org/10.1371/journal.pone.0238372.

Ranjbar-Sahraei, Bijan, Nees Jan van Eck, and Rutger de Jong. 2018. “Accuracy of Affiliation Information in Microsoft Academic: Implications for Institutional Level Research Evaluation.” In .

Robinson-Garcia, Nicolas, Rodrigo Costas, and Thed N. van Leeuwen. 2020. “Open Access Uptake by Universities Worldwide.” *PeerJ* 8 (July): e9410. https://doi.org/10.7717/peerj.9410.

Robinson-García, Nicolas, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 2016. “Analyzing Data Citation Practices Using the Data Citation Index.” *Journal of the Association for Information Science and Technology* 67 (12): 2964–75. https://doi.org/10.1002/asi.23529.

Rørstad, Kristoffer, and Dag W. Aksnes. 2015. “Publication Rate Expressed by Age, Gender and Academic Position – A Large-Scale Analysis of Norwegian Academic Staff.” *Journal of Informetrics* 9 (2): 317–33. https://doi.org/10.1016/j.joi.2015.02.003.

Ross-Hellauer, Tony, Stefan Reichmann, Nicki Lisa Cole, Angela Fessl, Thomas Klebel, and Nancy Pontika. 2021. “Dynamics of Cumulative Advantage and Threats to Equity in Open Science - A Scoping Review.” SocArXiv. https://doi.org/10.31235/osf.io/d5fz7.

Ryan, Jeffrey A., and Joshua M. Ulrich. 2020. *Quantmod: Quantitative Financial Modelling Framework* (version 0.4.18). https://CRAN.R-project.org/package=quantmod.

Siler, Kyle, and Koen Frenken. 2019. “The Pricing of Open Access Journals: Diverse Niches and Sources of Value in Academic Publishing.” *Quantitative Science Studies* 1 (1): 28–59. https://doi.org/10.1162/qss_a_00016.

Siler, Kyle, Stefanie Haustein, Elise Smith, Vincent Larivière, and Juan Pablo Alperin. 2018. “Authorial and Institutional Stratification in Open Access Publishing: The Case of Global Health Research.” *PeerJ* 6 (February): e4269. https://doi.org/10.7717/peerj.4269.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015a. “An Overview of Microsoft Academic Service (MAS) and Applications.” In *Proceedings of the 24th International Conference on World Wide Web - WWW ’15 Companion*, 243–46. Florence, Italy: ACM Press. https://doi.org/10.1145/2740908.2742839.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015b. “An Overview of Microsoft Academic Service (Mas) and Applications.” In *Proceedings of the 24th International Conference on World Wide Web*, 243–46.

Siriwardhana, Chesmal. 2015. “Promotion and Reporting of Research from Resource-Limited Settings.” *Infectious Diseases* 8: 25–29. https://doi.org/10.4137/IDRT.S16195.

Steinhardt, Isabel. 2020. “Learning Open Science by Doing Open Science. A Reflection of a Qualitative Research Project-Based Seminar.” *Education for Information* 36 (3): 263–79. https://doi.org/10.3233/EFI-190308.

Sweileh, Waleed M. 2020. “Bibliometric Analysis of Scientific Publications on ‘Sustainable Development Goals’ with Emphasis on ‘Good Health and Well-Being’ Goal (2015-2019).” *Globalization and Health* 16 (1): 68. https://doi.org/10.1186/s12992-020-00602-2.

Tennant, Jonathan P., François Waldner, Damien C. Jacques, Paola Masuzzo, Lauren B. Collister, and Chris. H. J. Hartgerink. 2016. “The Academic, Economic and Societal Impacts of Open Access: An Evidence-Based Review.” *F1000Research* 5 (September): 632. https://doi.org/10.12688/f1000research.8460.3.

Tenopir, Carol, Robert J. Sandusky, Suzie Allard, and Ben Birch. 2014. “Research Data Management Services in Academic Research Libraries and Perceptions of Librarians.” *Library & Information Science Research* 36 (2): 84–90. https://doi.org/10.1016/j.lisr.2013.11.003.

Tenopir, Carol, Sanna Talja, Wolfram Horstmann, Elina Late, Dane Hughes, Danielle Pollock, Birgit Schmidt, Lynn Baird, Robert Sandusky, and Suzie Allard. 2017. “Research Data Services in European

Academic Research Libraries." *LIBER Quarterly* 27 (1): 23–44. https://doi.org/10.18352/lq.10180.
Thomas, Emma G., Bamini Jayabalasingham, Tom Collins, Jeroen Geertzen, Chinh Bui, and Francesca Dominici. 2019. "Gender Disparities in Invited Commentary Authorship in 2459 Medical Journals." *JAMA Network Open* 2 (10): e1913682. https://doi.org/10.1001/jamanetworkopen.2019.13682.
Van Eck, Nees Jan. 2021. "CWTS Leiden Ranking 2020." Zenodo. https://doi.org/10.5281/zenodo.4745545.
Wais, Kamil. 2006. "Gender Prediction Methods Based on First Names with GenderizeR." *The R Journal* 8 (1): 17–37. https://doi.org/10.32614/RJ-2016-002.
———. 2016. "Gender Prediction Methods Based on First Names with GenderizeR." *The R Journal* 8 (1): 17–37. https://doi.org/10.32614/RJ-2016-002.
Wais, Kamil, Nathan VanHoudnos, John Ramey, and Thomas Klebel. 2019. *GenderizeR: Gender Prediction Based on First Names* (version 2.1.1). https://CRAN.R-project.org/package=genderizeR.
Waltman, Ludo. 2016. "A Review of the Literature on Citation Impact Indicators." *Journal of Informetrics* 10 (2): 365–91. https://doi.org/10.1016/j.joi.2016.02.007.
Waltman, Ludo, Clara Calero-Medina, Joost Kosten, Ed CM Noyons, Robert JW Tijssen, Nees Jan van Eck, Thed N van Leeuwen, Anthony FJ van Raan, Martijn S Visser, and Paul Wouters. 2012. "The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation." *Journal of the American Society for Information Science and Technology* 63 (12): 2419–32.
Waltman, Ludo, and Nees Jan van Eck. 2012. "A New Methodology for Constructing a Publication-Level Classification System of Science: A New Methodology for Constructing a Publication-Level Classification System of Science." *Journal of the American Society for Information Science and Technology* 63 (12): 2378–92. https://doi.org/10.1002/asi.22748.
———. 2019. "Field Normalization of Scientometric Indicators." In *Springer Handbook of Science and Technology Indicators*, edited by Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 281–300. Springer Handbooks. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_11.
Wang, Kuansan, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. "Microsoft Academic Graph: When Experts Are Not Enough." *Quantitative Science Studies* 1 (1): 396–413. https://doi.org/10.1162/qss_a_00021.
Wang, Kuansan, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. "A Review of Microsoft Academic Services for Science of Science Studies." *Frontiers in Big Data* 2 (December): 45. https://doi.org/10.3389/fdata.2019.00045.
Ware, Mark, and Michael Mabe. 2015. "The STM Report: An Overview of Scientific Andscholarly Journal Publishing." Fourth Edition. https://apo.org.au/sites/default/files/resource-files/2015-03/apo-nid57525.pdf.
West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. 2013. "The Role of Gender in Scholarly Authorship." *PLOS ONE* 8 (7): e66212. https://doi.org/10.1371/journal.pone.0066212.
Yair, Gad, and Keith Goldstein. 2020. "The Annus Mirabilis Paper: Years of Peak Productivity in Scientific Careers." *Scientometrics* 124 (2): 887–902. https://doi.org/10.1007/s11192-020-03544-z.
Zinkernagel, Roland, James Evans, and Lena Neij. 2018. "Applying the SDGs to Cities: Business as Usual or a New Dawn?" *Sustainability* 10 (9): 3201. https://doi.org/10.3390/su10093201.
Zuckerman, Harriet. 1988. "The Sociology of Science." In *Handbook of Sociology*, edited by Neil J. Smelser, 511–74. Thousand Oaks, CA, US: Sage Publications, Inc.
Zwart, Hub, Laurens Landeweerd, and Arjan van Rooij. 2014. "Adapt or Perish? Assessing the Recent Shift in the European Research Funding Arena from 'ELSA' to 'RRI.'" *Life Sciences, Society and Policy* 10 (1): 11. https://doi.org/10.1186/s40504-014-0011-x.
Zwart, Hub, and Annemiek Nelis. 2009. "What Is ELSA Genomics?" *EMBO Reports* 10 (6): 540–44. https://doi.org/10.1038/embor.2009.115.

10. Annex

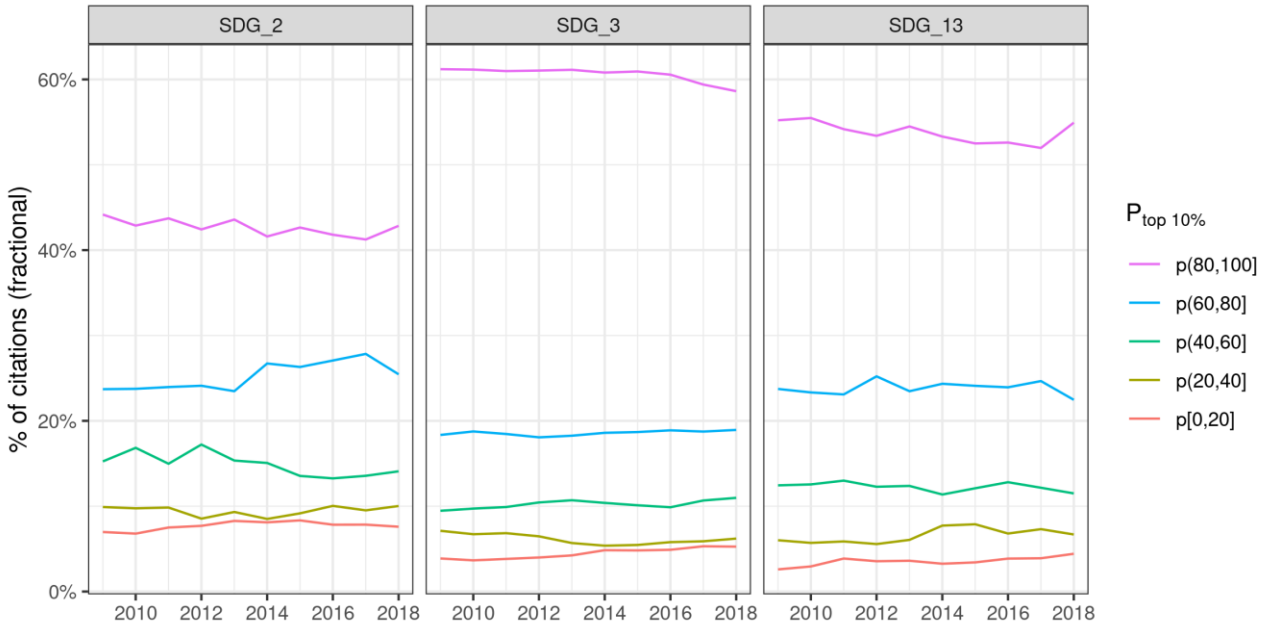

Figure A1: Share of fractional citations received by quintiles of the distribution of Ptop 10%

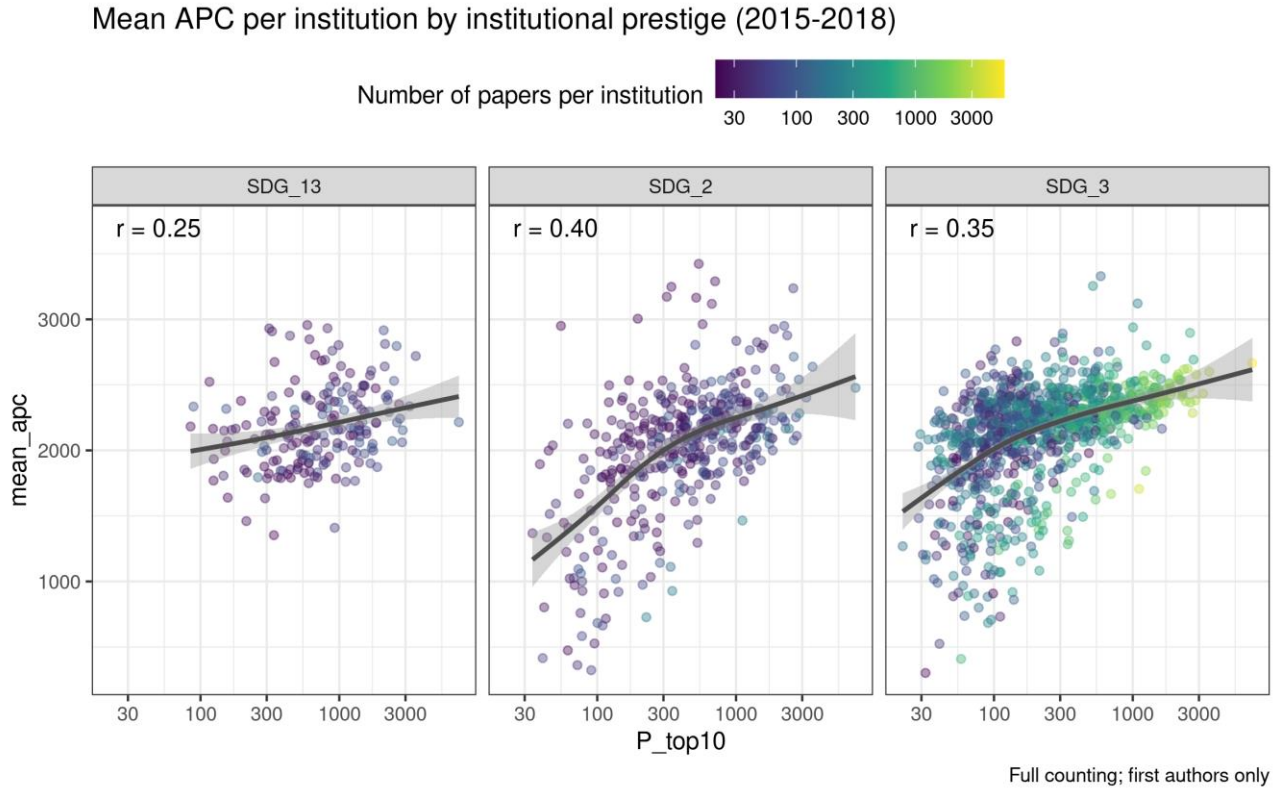

Figure A2: APC by institutional prestige without journals that do not have an APC; First authors

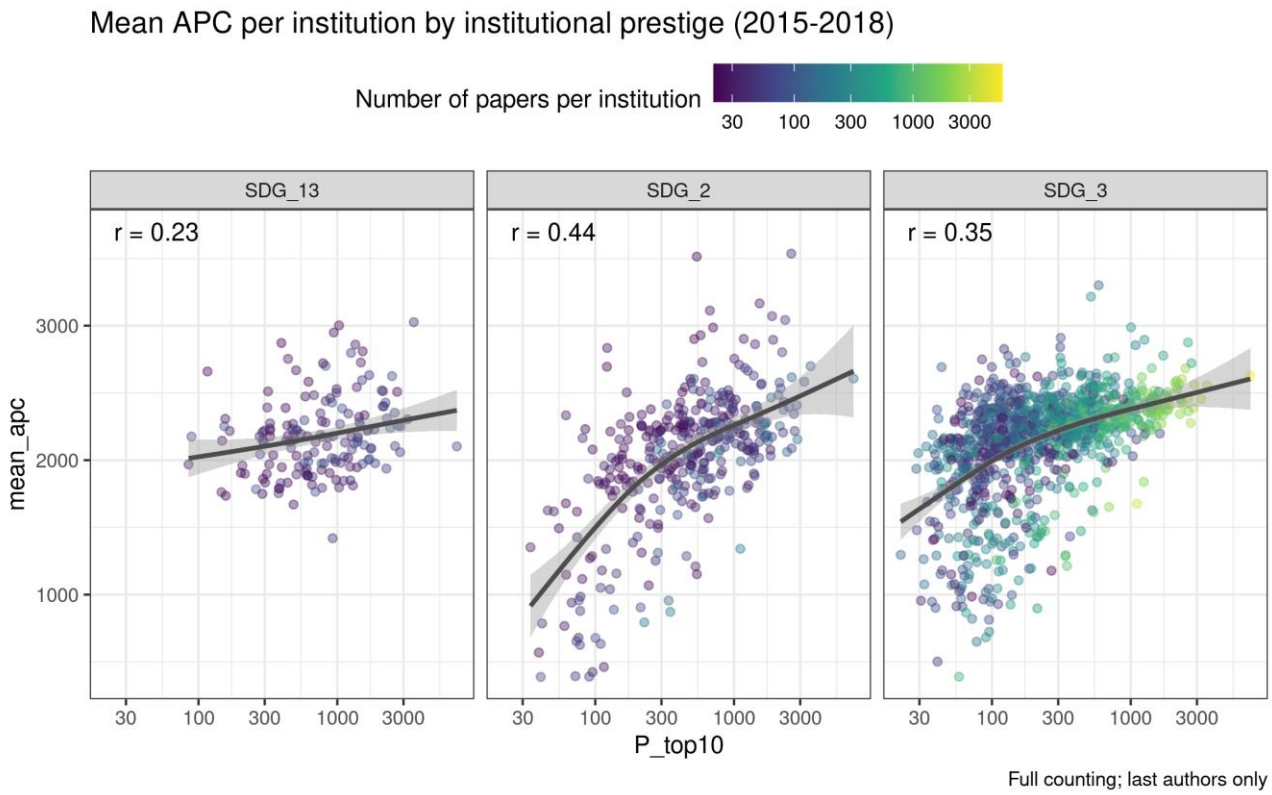

Figure A3: APC by institutional prestige without journals that do not have an APC; Last authors