Evaluating institutional open access performance: Sensitivity analysis

August 12, 2020

A companion white paper to the article "Evaluating institutional open access performance: Methodology, challenges and assessment"

Chun-Kai Huang[1], Cameron Neylon[1,2], Richard Hosking[2], Lucy Montgomery[1,2], Katie Wilson[1], Alkim Ozaygen[1], Chloe Brookes-Kenworthy[1]

[1]Centre for Culture and Technology, School of Media, Creative Arts and Social Inquiry, Curtin University, Kent St, Bentley 6102, Western Australia [2]Curtin Institute for Computation, Kent St, Bentley 6102, Western Australia

Abstract

In the article "Evaluating institutional open access performance: Methodology, challenges and assessment" we develop the first comprehensive and reproducible workflow that integrates multiple bibliographic data sources for evaluating institutional open access (OA) performance. The major data sources include Web of Science, Scopus, Microsoft Academic, and Unpaywall. However, each of these databases continues to update, both actively and retrospectively. This implies the results produced by the proposed process are potentially sensitive to both the choice of data source and the versions of them used. In addition, there remain the issue relating to selection bias in sample size and margin of error. The current work shows that the levels of sensitivity relating to the above issues can be significant at the institutional level. Hence, the transparency and clear documentation of the choices made on data sources (and their versions) and cut-off boundaries are vital for reproducibility and verifiability.

1 Introduction

Policy and implementation planning requires reliable, robust and relevant evidence. Much of the policy development for research resource allocation and strategy, including that focussed on the shift towards open access, relies on specific datasets, with known issues. Despite this, relatively few of the sources of evidence for research performance and evaluation test their sensitivity to the choices of data source and processing.

The evaluation of open access performance offers a useful case study of these issues. Early efforts to provide evidence on the extent of open access involved intensive manual data collection and analysis processes (Matsubayashi et al., 2009; Björk et al., 2010; Laakso et al., 2011). These do not lend themselves to regular updates or to tracking the effects of interventions. Much of the need for this manual work was driven by the challenges of discovering, identifying, disambiguating and determining the open access status of individual research outputs.

In general terms, analysing some aspect of research outputs, requires two sources of information. Firstly, data that connects outputs to the unit of analysis (e.g., organisation, funder, discipline). In our case we are interested in universities as a unit of analysis. Information linking research organisations to their outputs has traditionally been provided by the two major proprietary data providers, Web of Science and Scopus. More recently new players such as Digital Science have entered this space alongside Google Scholar and Microsoft Academic that both provide free services. Some affiliation data for research outputs is also provided by publishers through DOI registration agencies (particularly Crossref) but this is too patchy to be useful at this point.

The second source of information required is data linking individual outputs to the measure of interest. In our case this is open access status. The lack of a large-scale and comprehensive data source on open access status was the main reason driving the labour intensive and manual processes underpinning previous work. Over the past few years a number of services have emerged, with Unpaywall from OurResearch providing the most commonly used data.

A further critical issue is the means by which a set of outputs is unambigously connected to the data about those outputs in the second data source. This requires either that there be a shared unique identifier or a disambiguation process. As Unpaywall focuses on information about objects identified by Crossref DOIs we use Crossref as a source for attributes, such as publication date, that we require across all the outputs we examine. In other work we have shown that there are significant differences in the publication date recorded across different bibliographic data sources (Huang et al., 2020a).

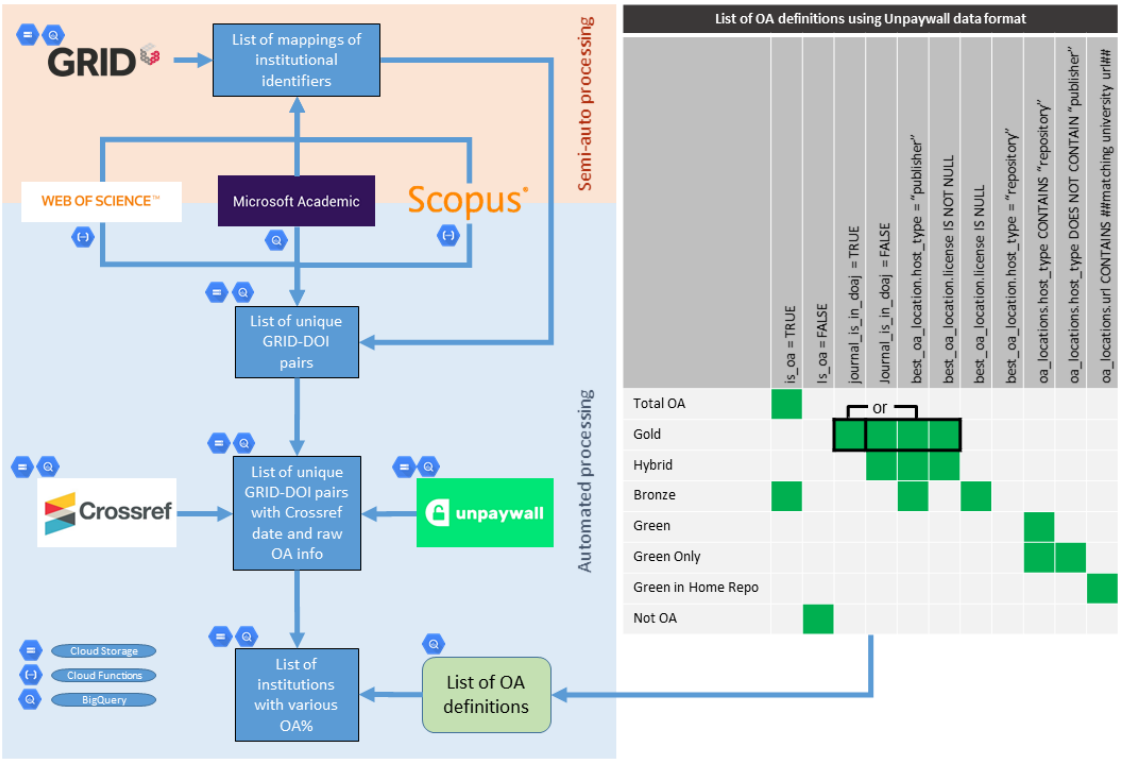
In Huang et al. (2020b) we propose and present the first comprehensive and reproducible workflow that integrates multiple data sources for evaluating university open access levels. However, more detailed study is required to understand the implications of the data workflow and how sensitive it is against various data sources and decisions made regarding the use of these data. This companion white paper aims to provide detailed analyses on how the results can be sensitive to the choice of data sources, and inclusion of universities due to level of confidence.

1.1 Fully specifying an analysis workflow

In Figure 1 we describe the generic workflow for analysis of research outputs described above and the specifics of how this applies to the analysis of open access performance by research organisations. To achieve full transparency the ideal would be a fully specified, and therefore reproducible, workflow with specific instances of input and output data identified. This is not straightforward.

There are significant tensions between maintaining up-to-date data across multiple sources while simultaneously providing granular detail on the specific versions of those data used as inputs. Data collection is neither instantaneous nor synchronous. For example, for Microsoft Academic we can obtain timestamped data dumps which can be uniquely identified. However, for Web of Science and Scopus we use API calls which take time to complete. There are additional limitations due to licensing conditions on our ability to fully share the data obtained from some proprietary sources.

Figure 1: Workflow of data collection and mapping of open access definitions to Unpaywall metadata.



These conditions mean that it is not possible, either in theory or in practice, to provide a completely replicable workflow that would allow a third party to obtain exactly the same results. We can work towards a goal of fully describing as much of the workflow as possible, and provide usable snapshots of derived data products from as early in the pipeline as possible, but improving on this will be an ongoing process.

Note here that, in contrast to some of the analysis performed in the main article (Huang et al., 2020b), no filtering of the list of universities is applied in this article. However, a detailed analysis of the margins of error and sample size is given in Section 5.

1.2 The importance of sensitivity analysis

Because any such data analysis pipeline will involve making choices that cannot be made fully transparent, it is crucial that we provide an analysis of how large an effect those choices have on the results that we present. This includes the choices of data sources, issues of the timing of data collection, the statistical properties of output data and the likely robustness of specific metrics.

The remainder of this article is structured as follows. Section 2 compares the use of Web of Science, Scopus, Microsoft Academic and the full combined dataset. The comparisons are made at the country level, region level and over time. Section 3 examines the effects of using different Unpaywall versions. Lastly, Section 4 explores the relationship between sample size and margin of error related to the estimation of percentages of various open access categories.

2 Sensitivity analysis on the use Web of Science, Scopus and Microsoft Academic

In this section we explore in detail how the uses of the three different bibliographic indexing databases (in particular, Web of Science, Scopus and Microsoft Academic) affect the resulting open access scores at the institutional level. For details on the data collection process and our definitions of open access scores, please see the main article Huang et al., (2020b). The analyses are grouped into three subsections that examines differences at country level, region level and over time (per year), respectively.

```
From combined
  …subtract scopus
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
  …subtract wos
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
  …subtract mag
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
```

```
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
From scopus
  …subtract wos
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
  …subtract mag
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
From wos
  …subtract scopus
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
```

```
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
  …subtract mag
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
From mag
  …subtract scopus
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
      …for value of percent_green
      …for value of percent_gold
      …for value of percent_hybrid
      …for value of percent_bronze
      …for value of percent_green_only
  …subtract wos
      …for value of total
      …for value of oa
      …for value of green
      …for value of gold
      …for value of hybrid
      …for value of bronze
      …for value of green_only
      …for value of green_in_home_repo
      …for value of percent_oa
```

```
…for value of percent_green
…for value of percent_gold
…for value of percent_hybrid
…for value of percent_bronze
…for value of percent_green_only
```

[10]:

	id	country	region	subregion	published_year	country_code	value	Combined Value	Source Value	comparison	difference_with	proxy
0	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2005	USA	1073.000000	2297.000000	1224.000000	Combined Dataset	Scopus	Total Publications Count
1	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2017	USA	2864.000000	5245.000000	2381.000000	Combined Dataset		
2	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2019	USA	3060.000000	3060.000000	0.000000	Combined Dataset		
3	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2010	USA	2159.000000	3897.000000	1738.000000	Combined Dataset		
4	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2018	USA	3450.000000	5553.000000	2103.000000	Combined Dataset		
…	…	…	…	…	…	…	…	…	…	…		
22520	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2009	CHN	-0.722123	9.840039	9.071856	Microsoft Academic		
22521	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2019	CHN	4.313352	4.310618	0.000000	Microsoft Academic		
22522	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2018	CHN	-0.730227	7.568964	7.821361	Microsoft Academic		
22523	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2003	CHN	1.047074	5.776173	4.240283	Microsoft Academic		
22524	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2006	CHN	1.299729	7.664955	5.561173	Microsoft Academic		

```
1              Scopus  Total Publications Count
2              Scopus  Total Publications Count
3              Scopus  Total Publications Count
4              Scopus  Total Publications Count
…               …                      …
22520  Web of Science         Green Only OA (%)
22521  Web of Science         Green Only OA (%)
22522  Web of Science         Green Only OA (%)
22523  Web of Science         Green Only OA (%)
22524  Web of Science         Green Only OA (%)

[2838150 rows x 12 columns]

['unpaywall_2019_11_22', 'unpaywall_2019_08_16', 'unpaywall_2019_04_19',
'unpaywall_2019_02_21', 'unpaywall_2018_09_24']

Comparing release unpaywall_2019_11_22
…with release unpaywall_2019_11_22
…with release unpaywall_2019_08_16
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_04_19
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
```

```
…with release unpaywall_2019_02_21
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2018_09_24
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
Comparing release unpaywall_2019_08_16
…with release unpaywall_2019_11_22
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_08_16
…with release unpaywall_2019_04_19
```

```
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_02_21
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2018_09_24
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
Comparing release unpaywall_2019_04_19
…with release unpaywall_2019_11_22
   …for proxy total
   …for proxy oa
```

```
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_08_16
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_04_19
…with release unpaywall_2019_02_21
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2018_09_24
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
```

```
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
Comparing release unpaywall_2019_02_21
…with release unpaywall_2019_11_22
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_08_16
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_04_19
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
```

```
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_02_21
…with release unpaywall_2018_09_24
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
Comparing release unpaywall_2018_09_24
…with release unpaywall_2019_11_22
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_08_16
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
```

```
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_04_19
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2019_02_21
   …for proxy total
   …for proxy oa
   …for proxy green
   …for proxy gold
   …for proxy hybrid
   …for proxy bronze
   …for proxy green_only
   …for proxy green_in_home_repo
   …for proxy percent_oa
   …for proxy percent_green
   …for proxy percent_gold
   …for proxy percent_hybrid
   …for proxy percent_bronze
   …for proxy percent_green_only
…with release unpaywall_2018_09_24
```

[14]:

	id	country	\
0	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
1	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
2	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
3	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
4	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
…	…	…	
22514	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	

```
22515  ffe0002dce72a38ae58852fe3e6f1b57                  Malaysia
22516  ffe0002dce72a38ae58852fe3e6f1b57                  Malaysia
22517  ffe0002dce72a38ae58852fe3e6f1b57                  Malaysia
22518  ffe0002dce72a38ae58852fe3e6f1b57                  Malaysia

              region           subregion  published_year country_code  \
0      North America    Northern America            2005          USA
1      North America    Northern America            2017          USA
2      North America    Northern America            2019          USA
3      North America    Northern America            2010          USA
4      North America    Northern America            2018          USA
…                …                  …               …           …
22514           Asia  South-eastern Asia            2015          MYS
22515           Asia  South-eastern Asia            2016          MYS
22516           Asia  South-eastern Asia            2017          MYS
22517           Asia  South-eastern Asia            2018          MYS
22518           Asia  South-eastern Asia            2019          MYS

       Later Value  Earlier Value  Difference         Later Release  \
0      2297.000000    2297.000000    0.000000  unpaywall_2019_11_22
1      5244.000000    5243.000000    1.000000  unpaywall_2019_11_22
2      3060.000000    2633.000000  427.000000  unpaywall_2019_11_22
3      3897.000000    3897.000000    0.000000  unpaywall_2019_11_22
4      5549.000000    5548.000000    1.000000  unpaywall_2019_11_22
…              …              …           …                   …
22514     2.923149       2.963311   -0.040162  unpaywall_2018_09_24
22515     2.066698       2.578528   -0.511830  unpaywall_2018_09_24
22516     2.078138       2.405641   -0.327503  unpaywall_2018_09_24
22517     1.092299       2.355372   -1.263073  unpaywall_2018_09_24
22518     1.923077       0.544959    1.378118  unpaywall_2018_09_24

            Earlier Release                     proxy
0      unpaywall_2019_08_16  Total Publications Count
1      unpaywall_2019_08_16  Total Publications Count
2      unpaywall_2019_08_16  Total Publications Count
3      unpaywall_2019_08_16  Total Publications Count
4      unpaywall_2019_08_16  Total Publications Count
…                     …                         …
22514  unpaywall_2019_02_21         Green Only OA (%)
22515  unpaywall_2019_02_21         Green Only OA (%)
22516  unpaywall_2019_02_21         Green Only OA (%)
22517  unpaywall_2019_02_21         Green Only OA (%)
22518  unpaywall_2019_02_21         Green Only OA (%)

[6306664 rows x 12 columns]
```

2.1 Comparisons under groupings by country

Firstly, we reproduce the distributions of institutional OA scores (in terms of Total OA, Gold OA and Green OA) as in the main article (Huang et al., 2020b), grouped by country. This is labelled as Figure 2. Subsequently, we show the same display but restrict the underlying data to each of Microsoft Academic, Web of Science, and Scopus, respectively (Figures 3 to 5). In each of these figures, each dot represent the OA score of an university from the specific country, for 2017. The colour code for regions (as indicated in Figure 2) will be used throughout this article, where applicable.

Figure 2: With the combined full dataset - 2017 percentages of Total OA, Gold OA and Green OA (left to right) grouped by countries.

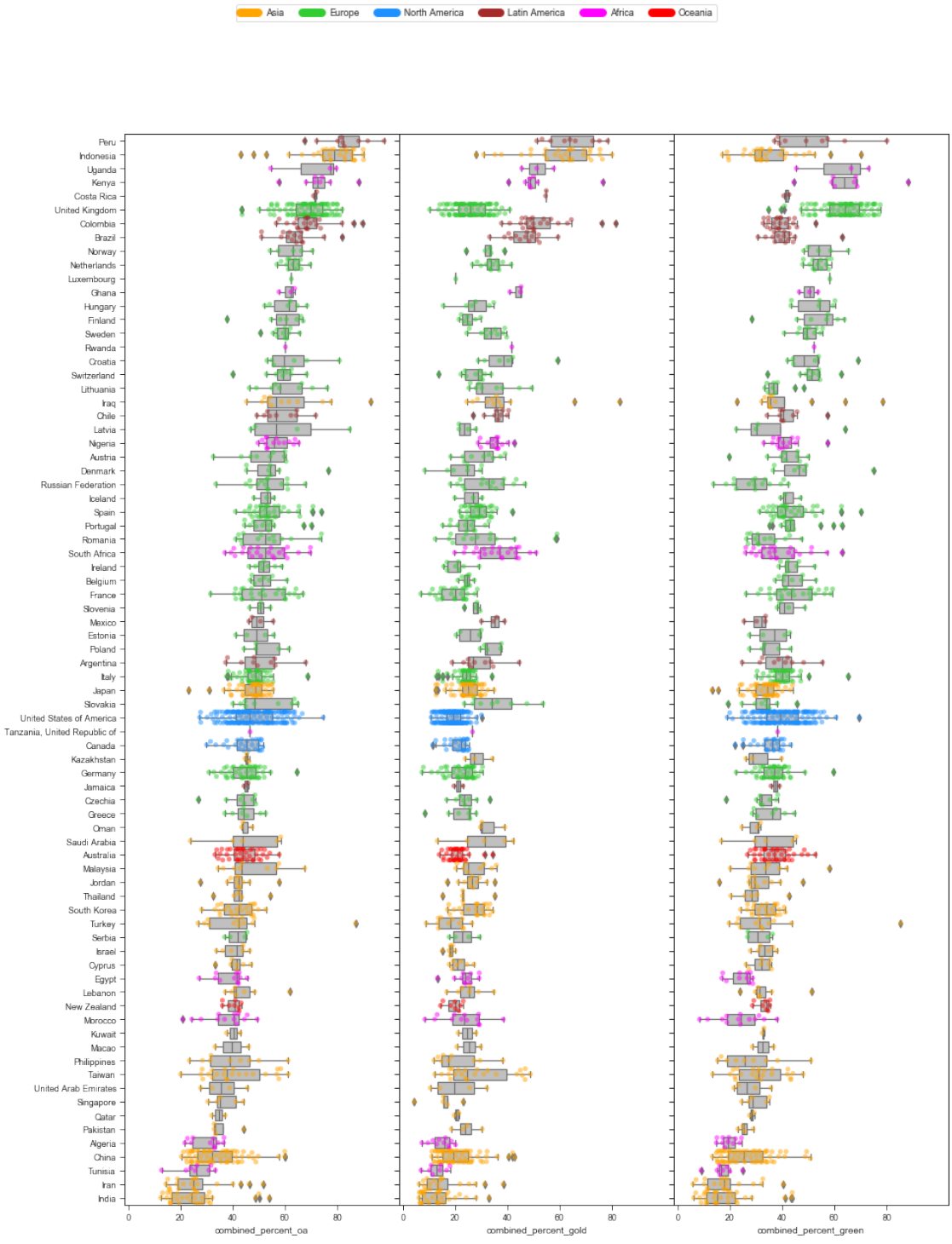


Figure 3: With only the Microsoft Academic dataset - 2017 percentages of Total OA, Gold OA and Green OA (left to right) grouped by countries.

Figure 4: With only the Web of Science dataset - 2017 percentages of Total OA, Gold OA and Green OA (left to right) grouped by countries.

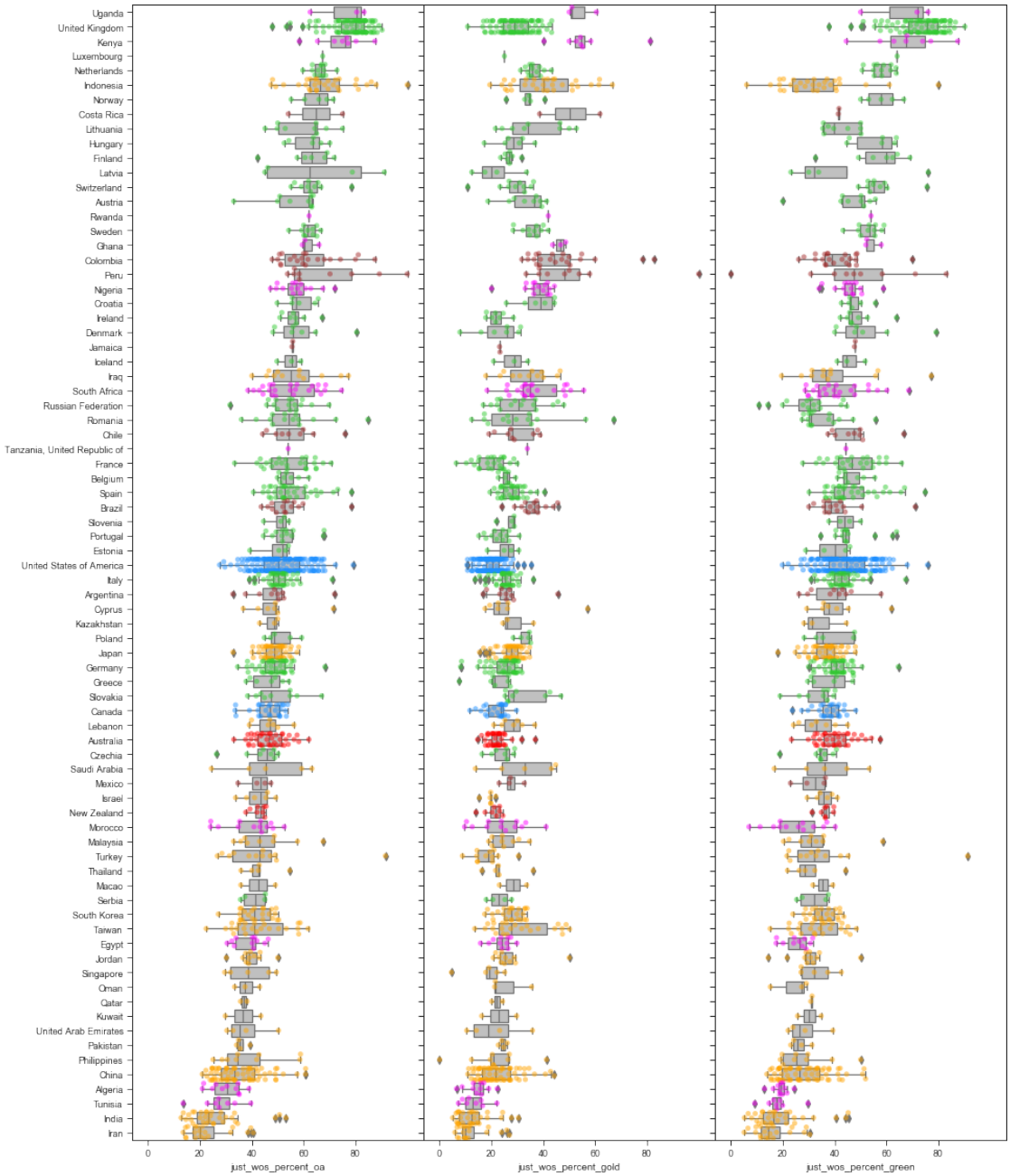


Figure 5: With only the Scopus dataset - 2017 percentages of Total OA, Gold OA and Green OA (left to right) grouped by countries.

In comparing Figures 2 to 5, we first note changes in the ordering of countries. In each figure, the order is determined by the median percentage of total open access (first column). These may be caused by the differences in coverage, as noted in Huang et al. (2020a). It is interesting to note the shifts of open access scores due to the choices of sources vary across countries. Not only do the median country open access scores shifts, the shapes of distributions of the various open access scores also change across datasets.

One of the most noticeable changes are that of the UK universities. The open access scores for UK universities remain mostly unchanged when the data is shifted from the full combined set to the one that only include Microsoft Academic records. However, the Green open access percentages seem to consistently shift upwards when using Web of Science or Scopus records only. This also results in higher total open access percentages for UK universities when only Web of Science or Scopus is used as the basis for DOI extraction.

To make the country level differences more clear, we now display the same graphs but for the open access percentage differences in Figures 6 to 11. Figures 6 to 8 shows the differences between the full dataset against each of Microsoft Academic, Web of Science, and Scopus. Figures 9 to 11 presents the differences amongst pairs of the three sources.

[19]:

	id	country	region	subregion	published_year	country_code	value	Combined Value	Source Value	comparison
1	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2017	USA	0.706245	61.944709	61.238464	Combined Dataset
24	3f3dac00919266d9156e20abf3e05d0a	Netherlands	Europe	Western Europe	2017	NLD	-0.092723	65.906906	65.999629	Combined Dataset
44	293018e797a8353364a328b1af4518b5	United States of America	North America	Northern America	2017	USA	-1.209123	38.852989	40.062112	Combined Dataset
70	4cc38006c47b39f0e8dbfa054ea44943	Ghana	Africa	Sub-Saharan Africa	2017	GHA	0.856237	57.674419	56.818182	Combined Dataset
78	0fd955e857845e0179b66816ae907e17	Argentina	Latin America	Latin America and the Caribbean	2017	ARG	2.138623	37.987680	35.849057	Combined Dataset
…	…	…	…	…	…	…	…	…	…	…
22441	9aaa25d74d1694a0df5a1e3ad79d41bf	Germany	Europe	Western Europe	2017	DEU	11.439298	40.978593	29.539295	Combined Dataset
22455	5a608e2f933f8c5e9aa494d5ee000607	United States of America	North America	Northern America	2017	USA	4.354647	42.661996	38.307350	Combined Dataset
22479	3d18990445717f490a7037ce0c43b223	United States of America	North America	Northern America	2017	USA	2.741710	41.350981	38.609272	Combined Dataset
22492	92f84dd0c88bb4c0823383f8026e0b37	South Korea	Asia	Eastern Asia	2017					
22506	8fac4241d860e90d4eced37eae996fd5	China	Asia	Eastern Asia	2017					

```
22492          KOR   7.148407       47.100176     39.951768  Combined Dataset
22506          CHN   6.948852       41.894893     34.946041  Combined Dataset

          difference_with            proxy
1      Microsoft Academic  Open Access (%)
24     Microsoft Academic  Open Access (%)
44     Microsoft Academic  Open Access (%)
70     Microsoft Academic  Open Access (%)
78     Microsoft Academic  Open Access (%)
…                    …            …
22441  Microsoft Academic  Open Access (%)
22455  Microsoft Academic  Open Access (%)
22479  Microsoft Academic  Open Access (%)
22492  Microsoft Academic  Open Access (%)
22506  Microsoft Academic  Open Access (%)

[1206 rows x 12 columns]
```

Figure 6: Differences in 2017 OA percentages for each institution between the full combined dataset and Microsoft Academic, grouped by country.

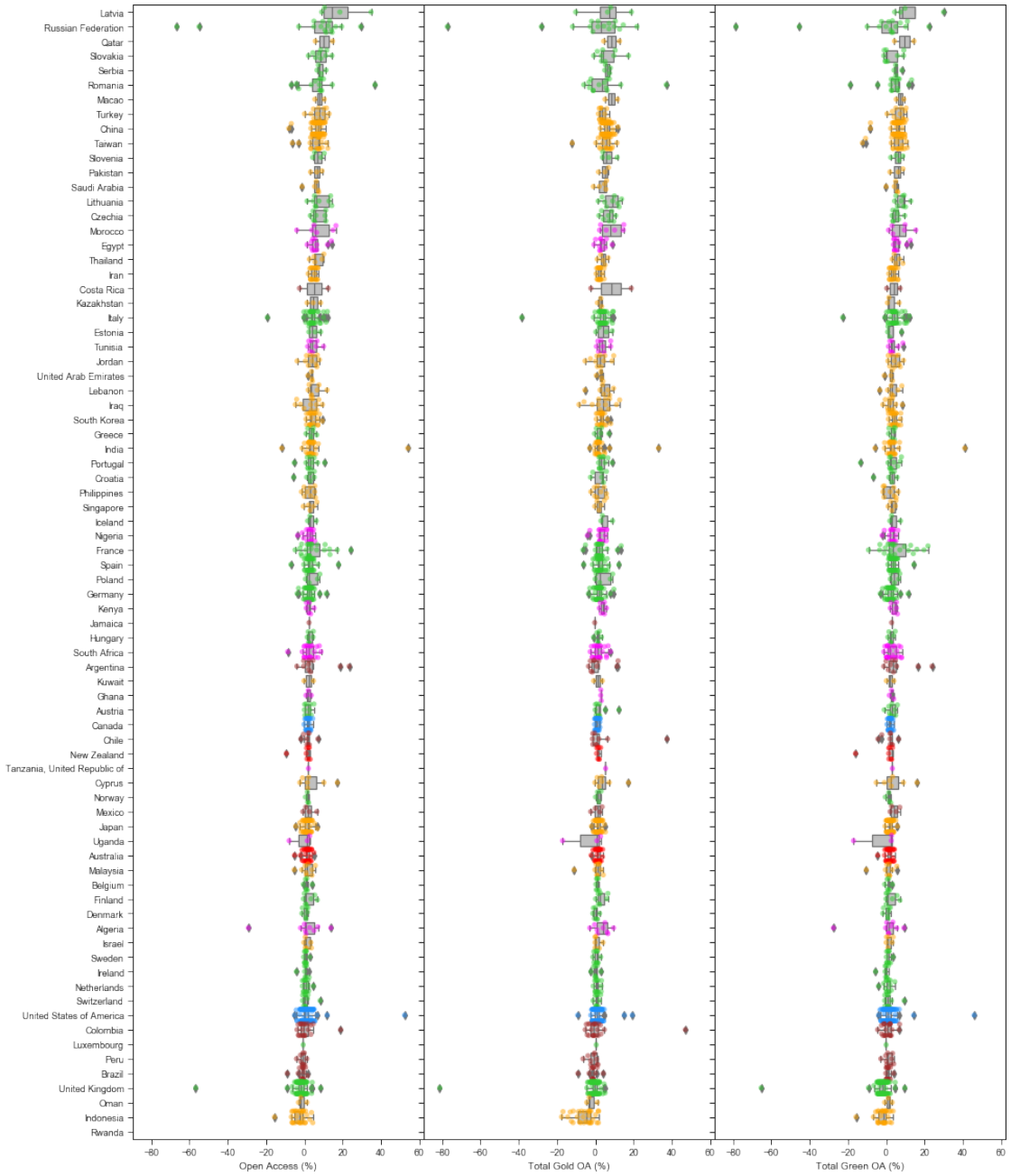


Figure 7: Differences in 2017 OA percentages for each institution between the full combined dataset and Web of Science, grouped by country.

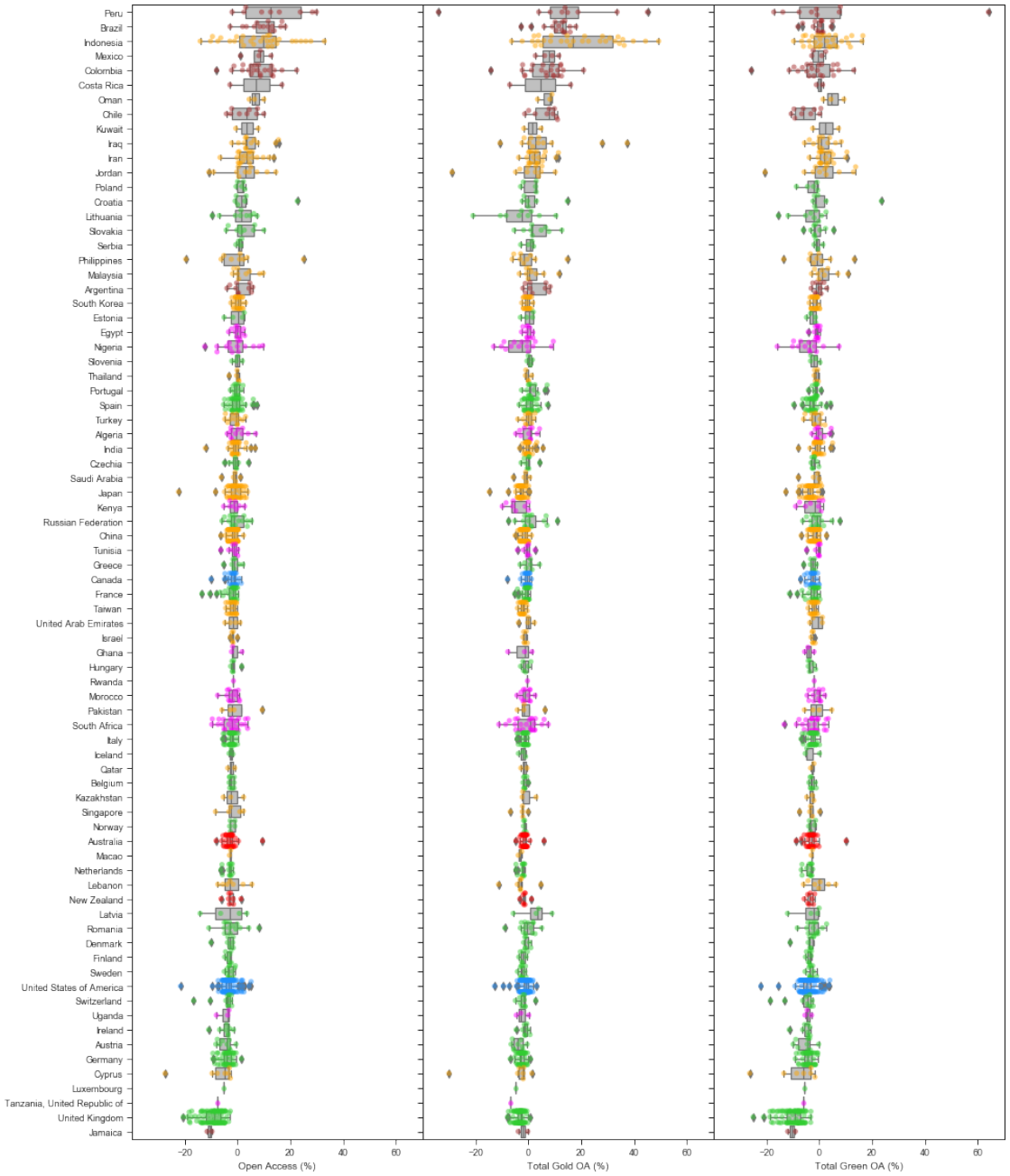


Figure 8: Differences in 2017 OA percentages for each institution between the full combined dataset and Scopus, grouped by country.

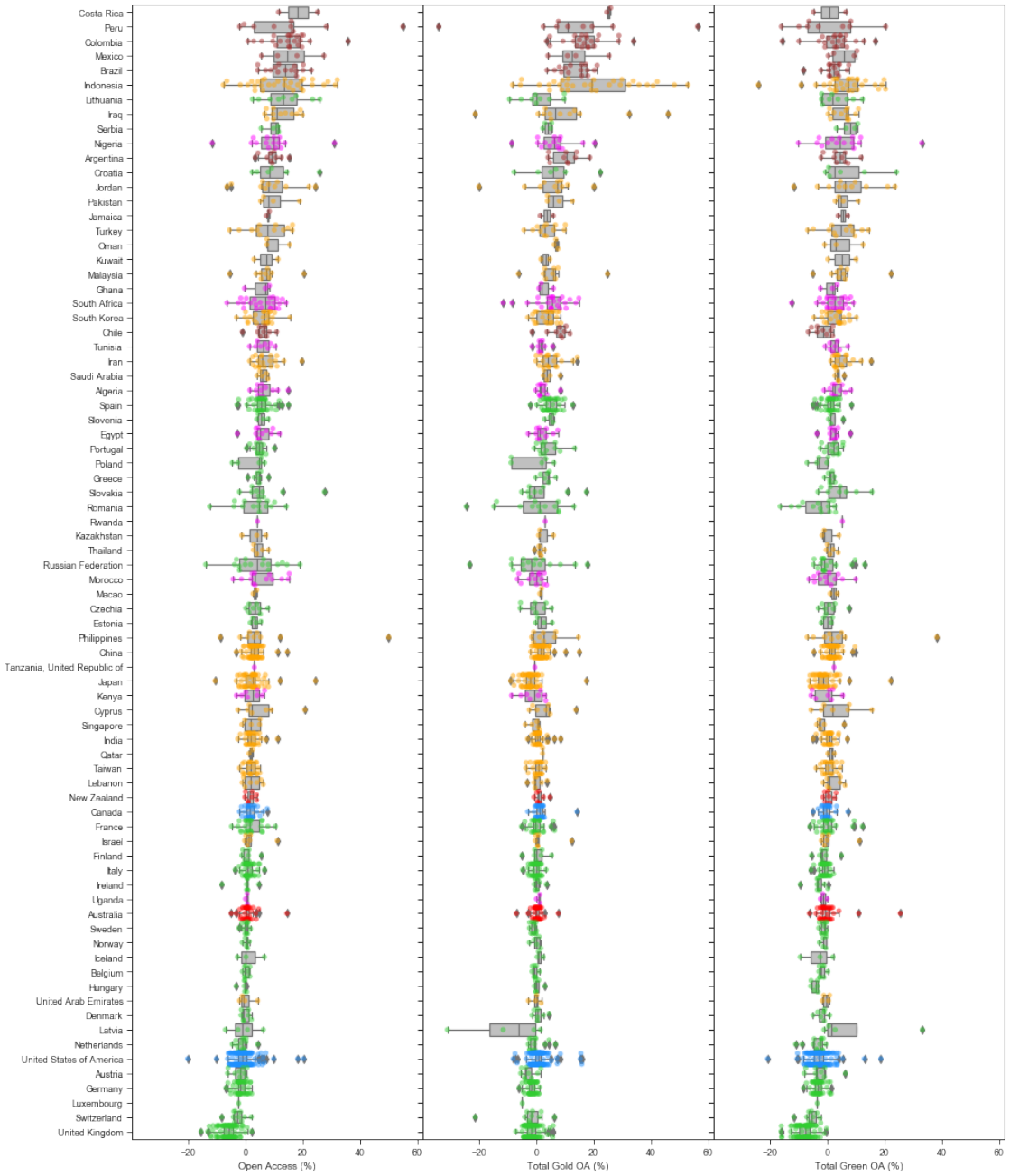


Figure 9: Differences in 2017 OA percentages for each institution between Web of Science and Scopus, grouped by country.

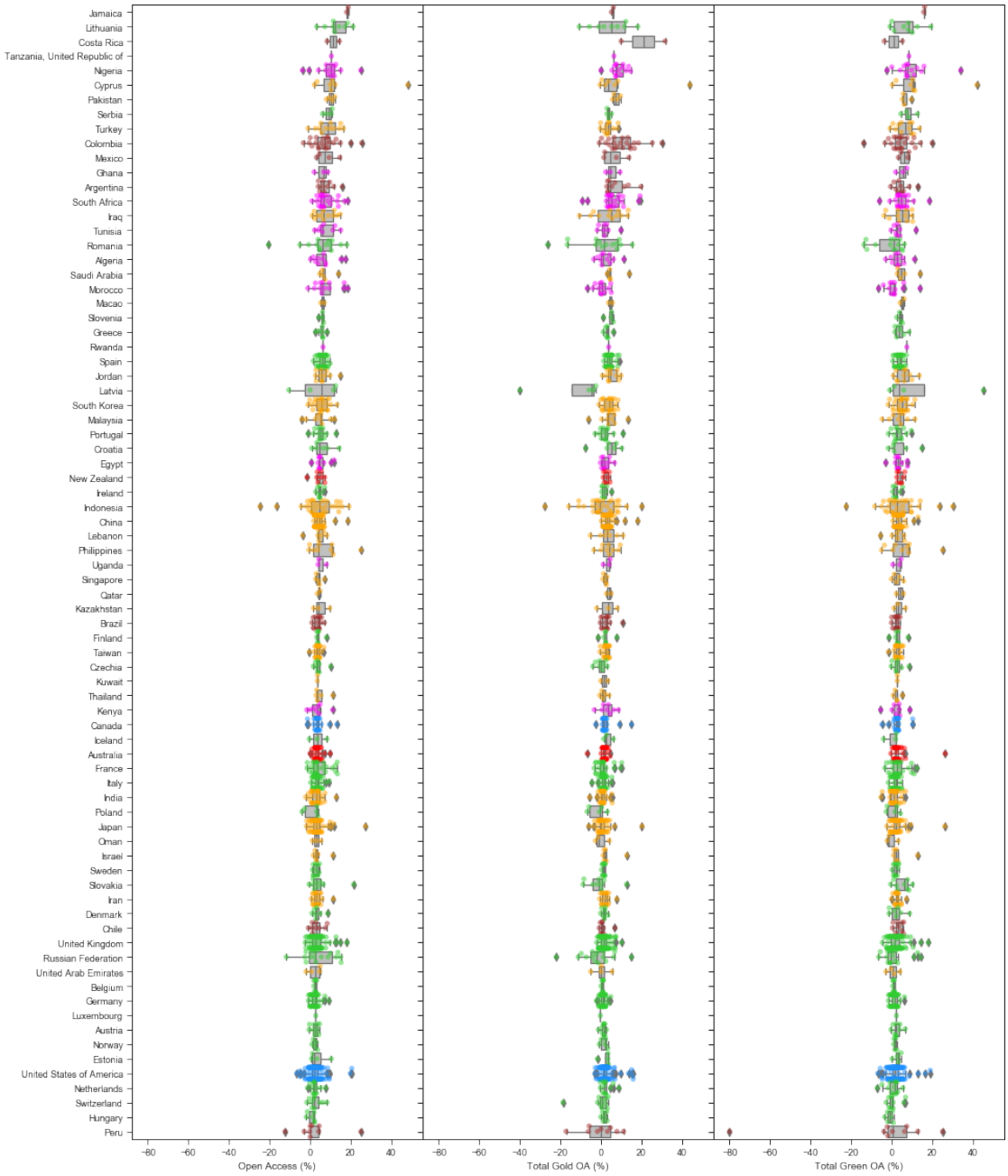
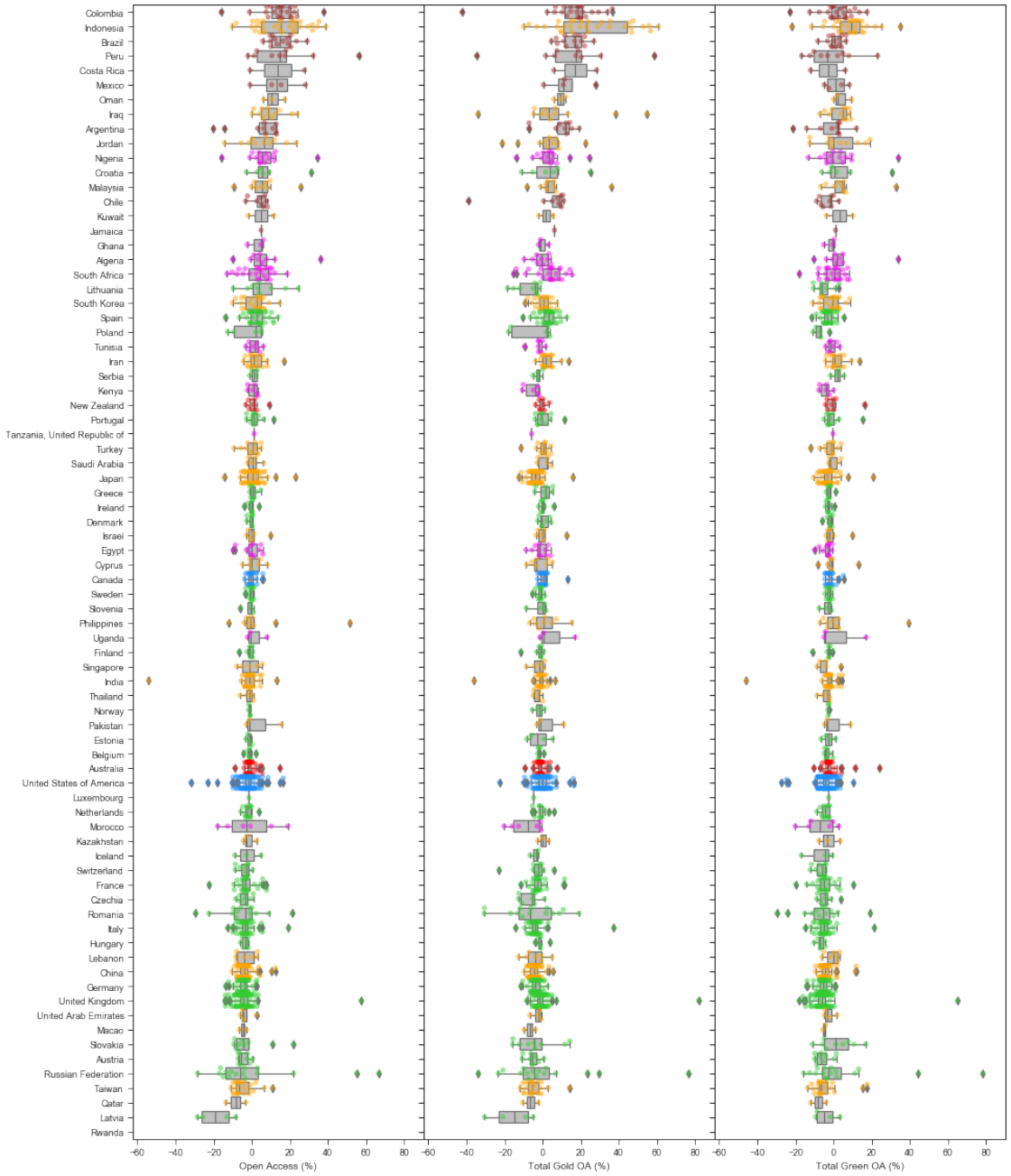


Figure 10: Differences in 2017 OA percentages for each institution between Web of Science and Microsoft Academic, grouped by country.

Figure 11: Differences in 2017 OA percentages for each institution between Microsoft Academic and Scopus, grouped by country.



In each of the Figures 6 to 11, the countries are ordered by the median differences in total open access percentages. For total open access, the shifts are mainly positive for universities in Brazil, Indonesia and Mexico, when the full combined dataset is used instead of Web of Science or Scopus only. The detailed effects on gold open access and green open access levels differ from that of total open acess. In particular, these universities' gold open access percentages are higher (positive shifts), and green open access percentages lower (negative shifts), for the full combined dataset. In

contrast, there are negative shifts for all three categories of open access relative to Web of Science and Scopus. The parallel comparison of the full combined dataset against Microsoft Academic shows substantially smaller differences, with most countries depicting a median difference around zero.

The comparisons between the three data sources show relatively less differences in results between Web of Science and Scopus, when contrasted against their differences in results against Microsoft Academic. These are indications that Microsoft Academic potentially have more comprehensive coverage of research outputs, especially for a select number of countries. This is consistent with what we have previously observed (Huang et al., 2020a).

2.2 Comparisons under groupings by region

As an alternative view of the above, we next group the universities into their respective regions and compares the resulting differences across these regions. Results are again for the year 2017. Figure 12 presents the various boxplots of the differences in total open access percentages (row 1), gold open access percentages (row 2) and green open access percentages (row 3) when contrasting the full combined dataset against each of Web of Science (column 1), Scopus (column 2) and Microsoft Academic (column 3). Figure 13 displays parallel visualisations for comparing open access percentage differences across any pairs of Web of Science, Scopus and Microsoft Academic.

Figure 12: Differences in OA% when contrasting the full combined dataset again each of Web of Science, Scopus and Microsoft Academic.

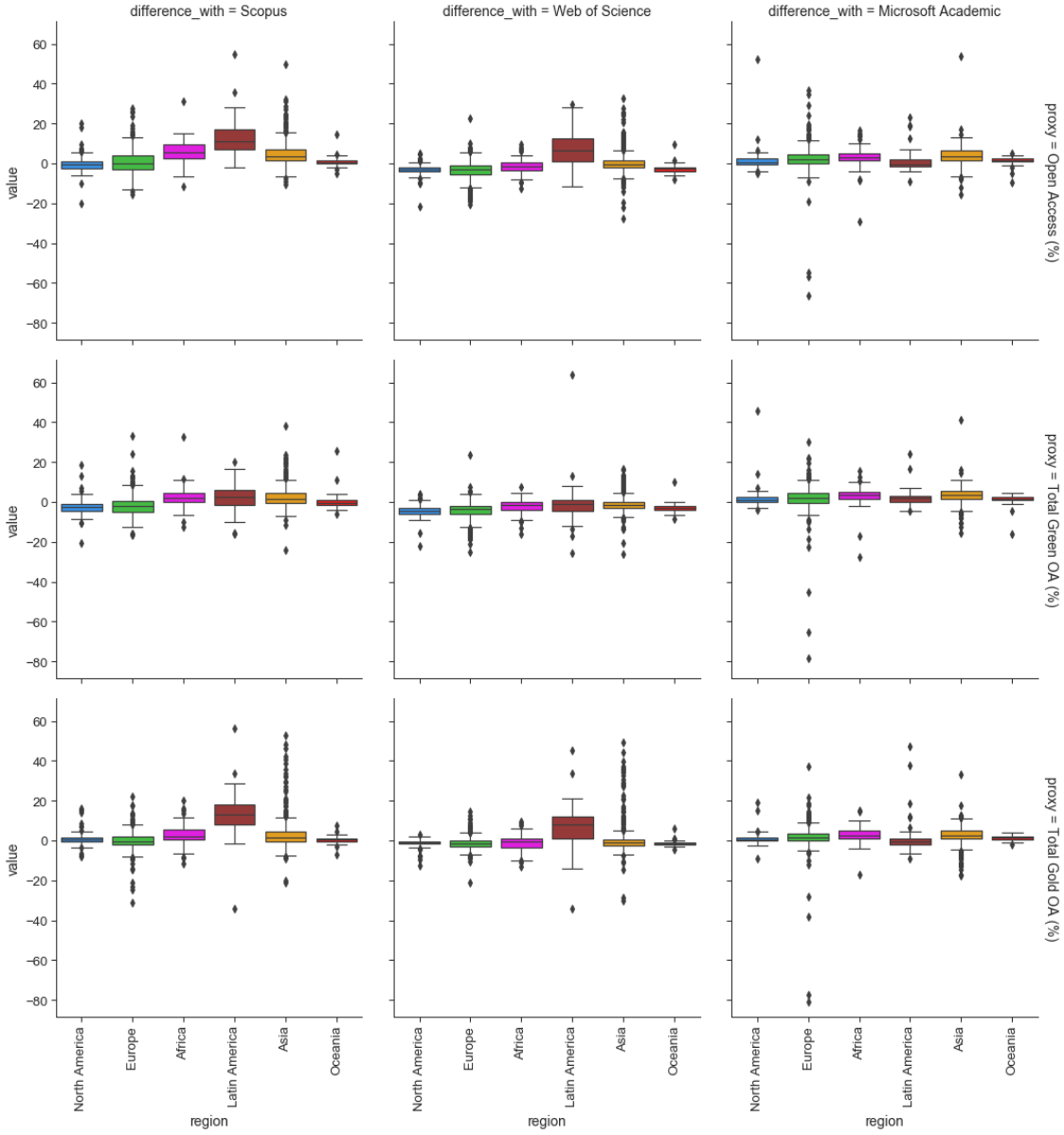
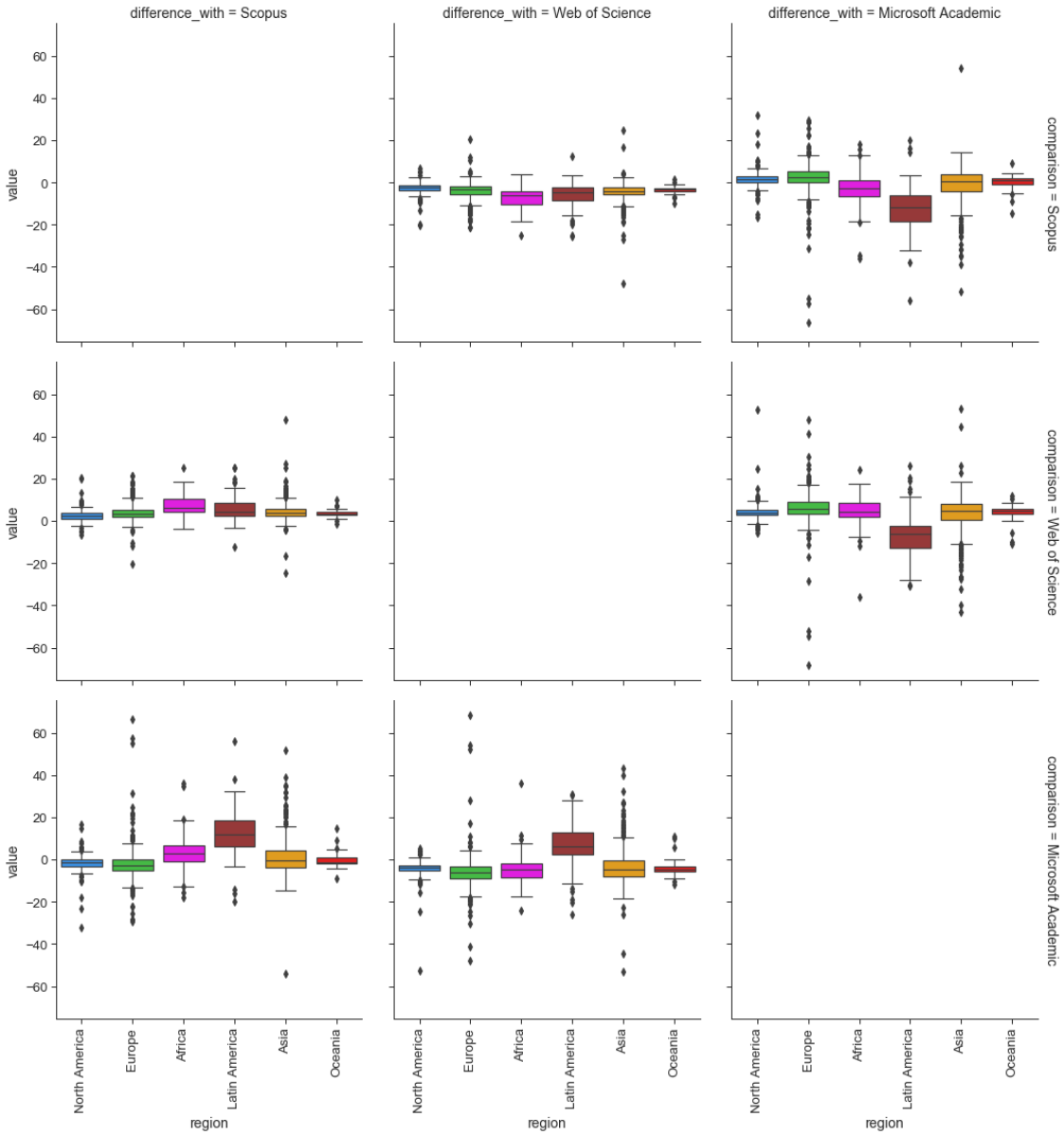


Figure 13: Differences in OA% when contrasting pairs of Web of Science, Scopus and Microsoft Academic.



The figures clearly depicts an advantage for Latin American universities when the full combined dataset, or Microsoft Academic, is used as opposed to using only Webs of Science or Scopus. The indication that Microsoft Academic produces a similar overall result to the full combined dataset is also re-emphasised in these graphs. However there remain large differences for specific universities. It is also worth noting the high number of outliers for Asian universities when contrasting the the combined set against Web of Science and Scopus, and similarly when comparing across pairs of the three data sources.

Combined with findings in the previous section, we have noted that there are different effects of

including Microsoft Academic depending on the open access categories and regions. Latin America sees higher proportions of both total open access and gold open access, while cases of lower green open access are also observed. This highlights the region's stronger focus on open access publishing rather than repository-mediated access to contents. In contrast, UK universities see lower total open access and green open access levels when Microsoft Academic is included. This is potentially attributed to the data contributing research outputs that are beyond the English language and European venues, which are less likely to have strong focuses on repository-mediated access, or where open access via repositories are less likely to be captured by the current systems.

2.3 Comparisons of differences across years

This subsection explores how the differences in various OA% across different pairs of datasets vary across years. Readers should note that the abnormality observed for the year 2019 is mainly due to the time delay in data collection, particularly from Web of Science and Scopus. We leave 2019 in the visualisation to illustrate some of the limitations of our data pipeline. Figure 14 describes the differences in total open access percentages between the combined data set and each of the individual data sources. Corresponding figures for gold open access and green open access are presented in Figures 15 and 16.

The median difference across all boxplots is close to zero. However, a significant number of boxplots for the differences between the combined dataset against Web of Science and Scopus are characterised by positive skewness. This is particularly evidenced for the gold open access percentages, and followed by the total open access percentages. This implies a significant portion of the universities are assigned higher gold open access percentages by the combined dataset, as compared to using only Web of Science or Scopus. This seems to then result in higher total open access percentages as well. In contrast, the boxplots related to Microsoft Academic seem more symmetric and also seem to have a smaller spread (in terms of the interquartile range).

No clear trend is observed for comparisons across time. Except, there appears to be a slight increase in discrepancies (in terms of the total open access percentages and green open access percentages) as we move further back in time (evidenced by the increases in range and interquartile range).

Figure 14: Difference in Total OA% over time between the full combined dataset against each of Web of Science, Scopus and Microsoft Academic.

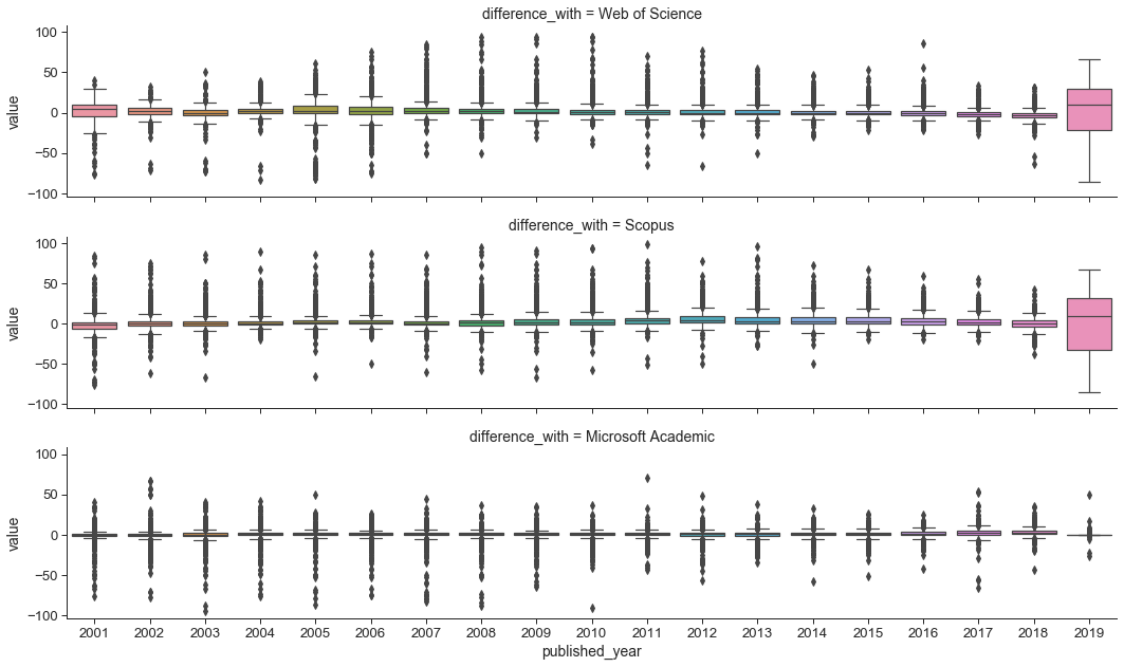

difference_with = Web of Science
difference_with = Scopus
difference_with = Microsoft Academic
value
100
50
0
−50
−100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
published_year

Figure 15: Difference in Gold OA% over time between the full combined dataset against each of Web of Science, Scopus and Microsoft Academic.

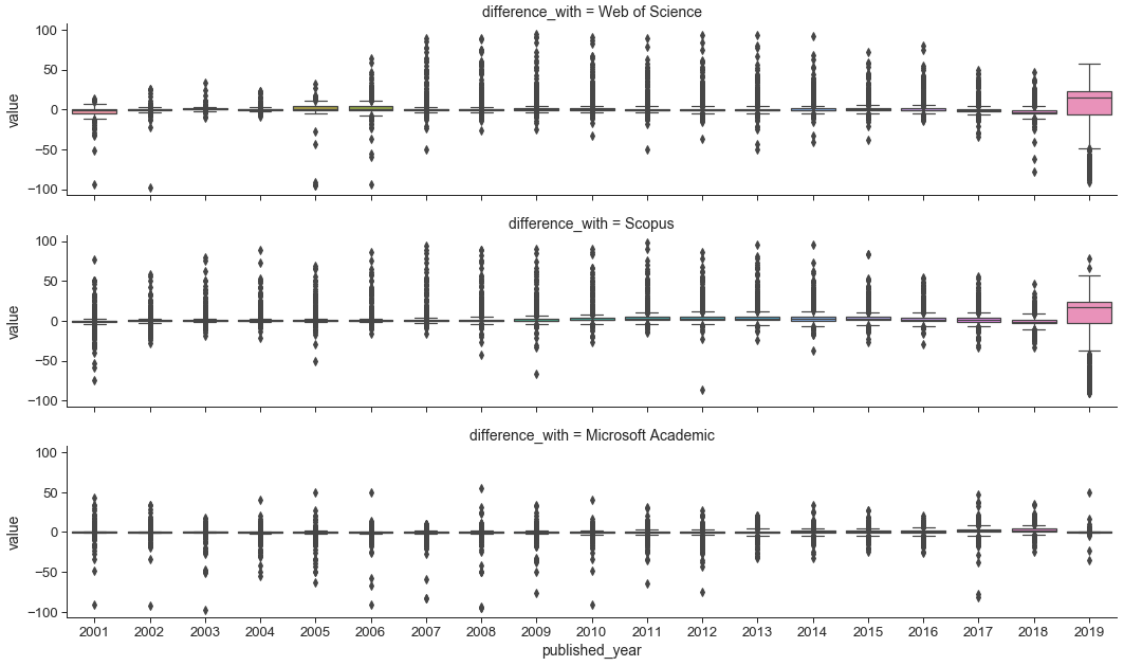


Figure 16: Difference in Green OA % over time between the full combined dataset against each of Web of Science, Scopus and Microsoft Academic.

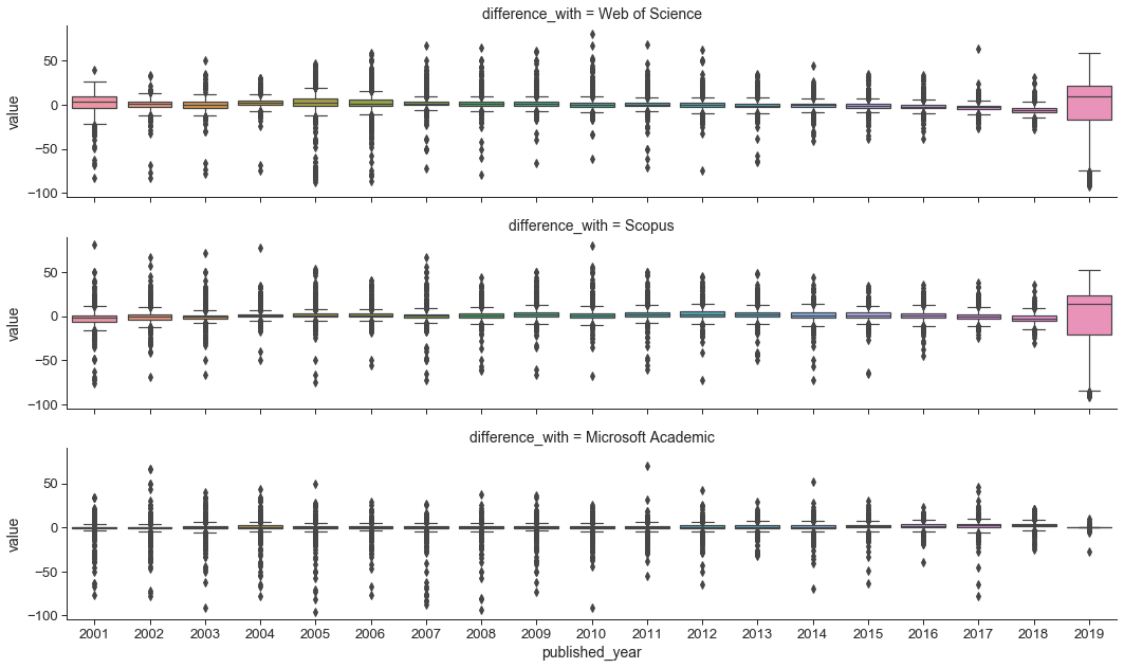
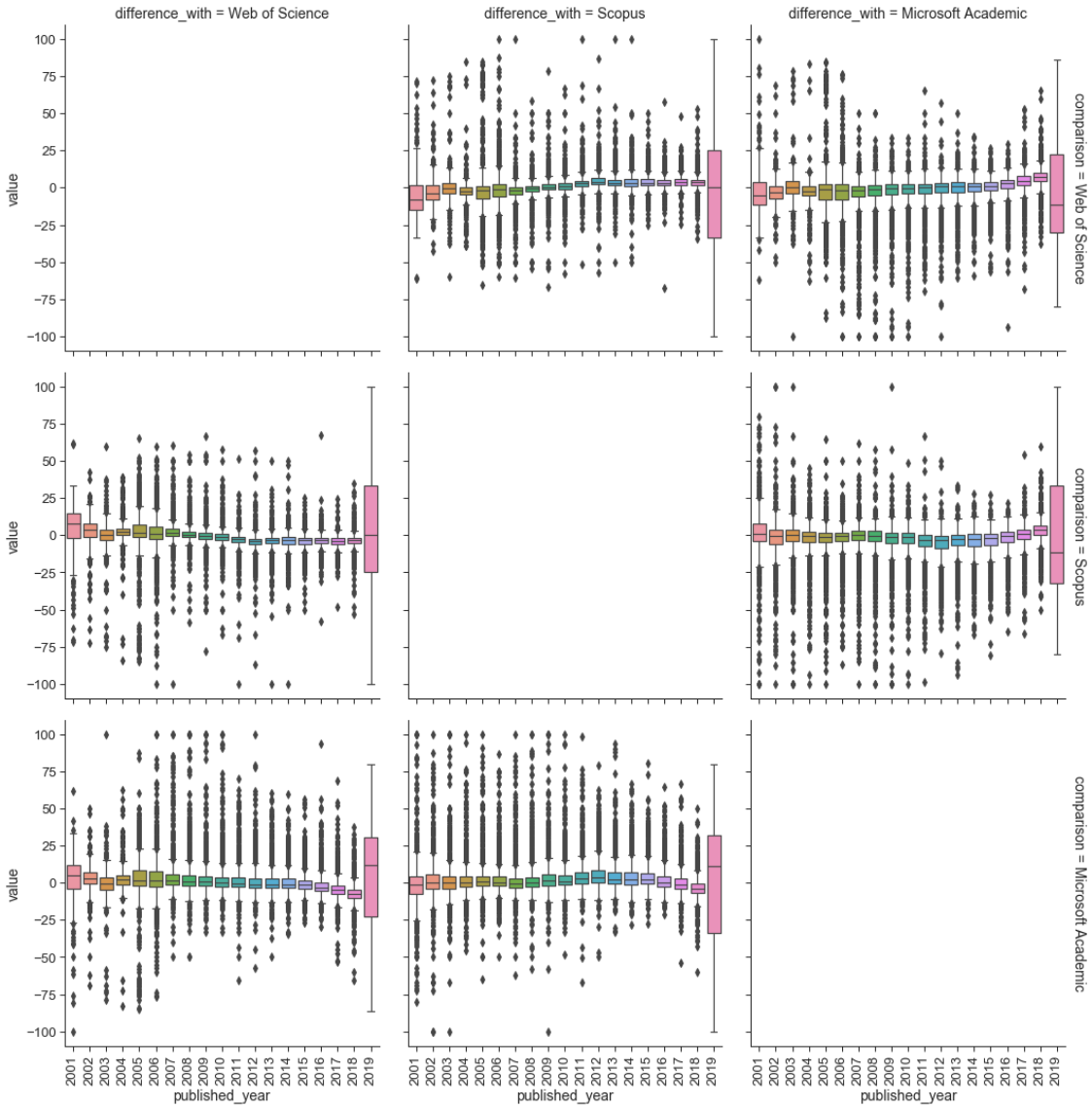
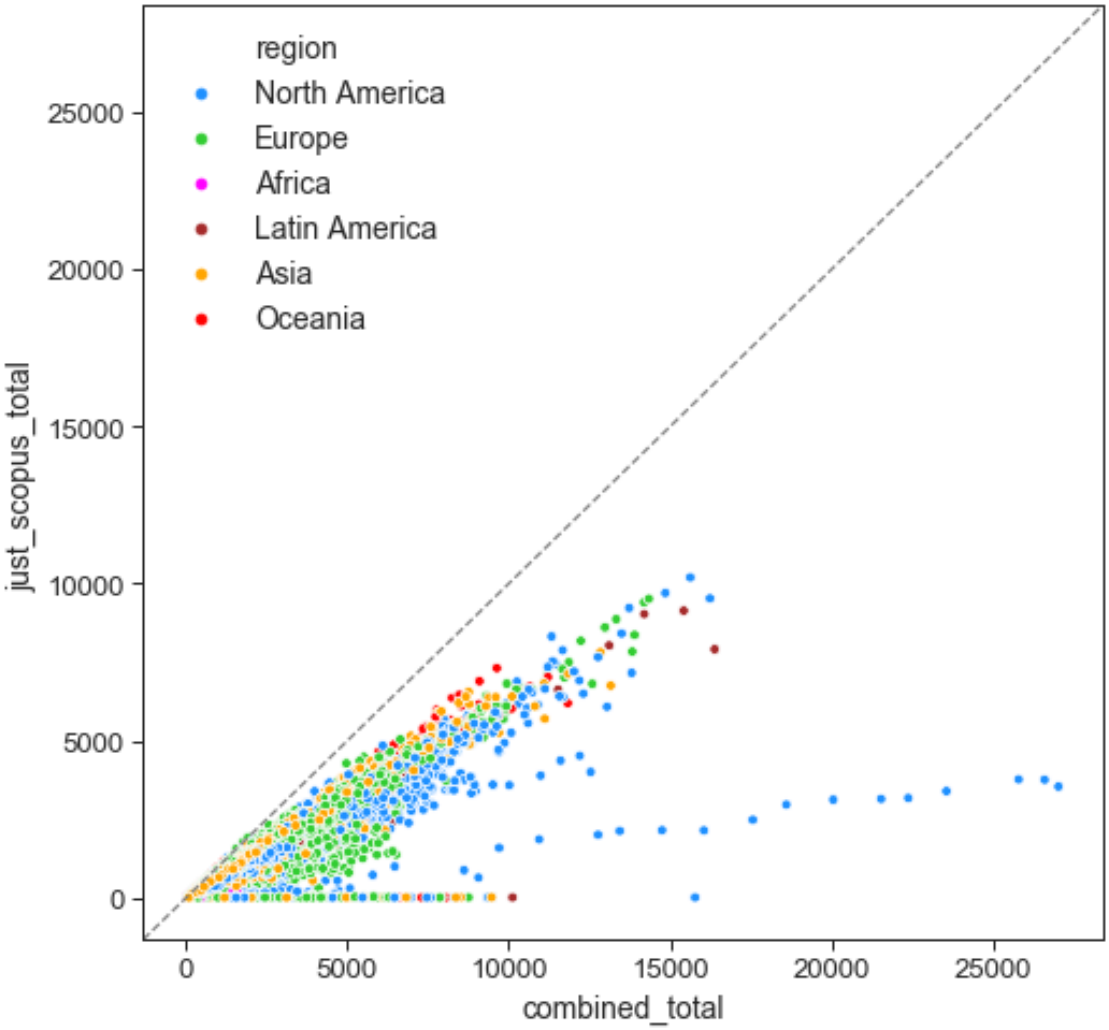


Figure 17 presents the corresponidng comparisons of total open access percentages over time for pairs of Web of Science, Scopus and Microsoft Academic. Many of the boxplots for the difference between Microsoft Academic against the other two sources depicted positive skewness. There is again signals of very slight increase of differences further back in time.

Figure 17: Difference in total OA% over time between any pairs of Web of Science, Scopus and Microsoft Academic.




region
North America
Europe
Africa
Latin America
Asia
Oceania
just_scopus_total
combined_total
0
5000
10000
15000
20000
25000

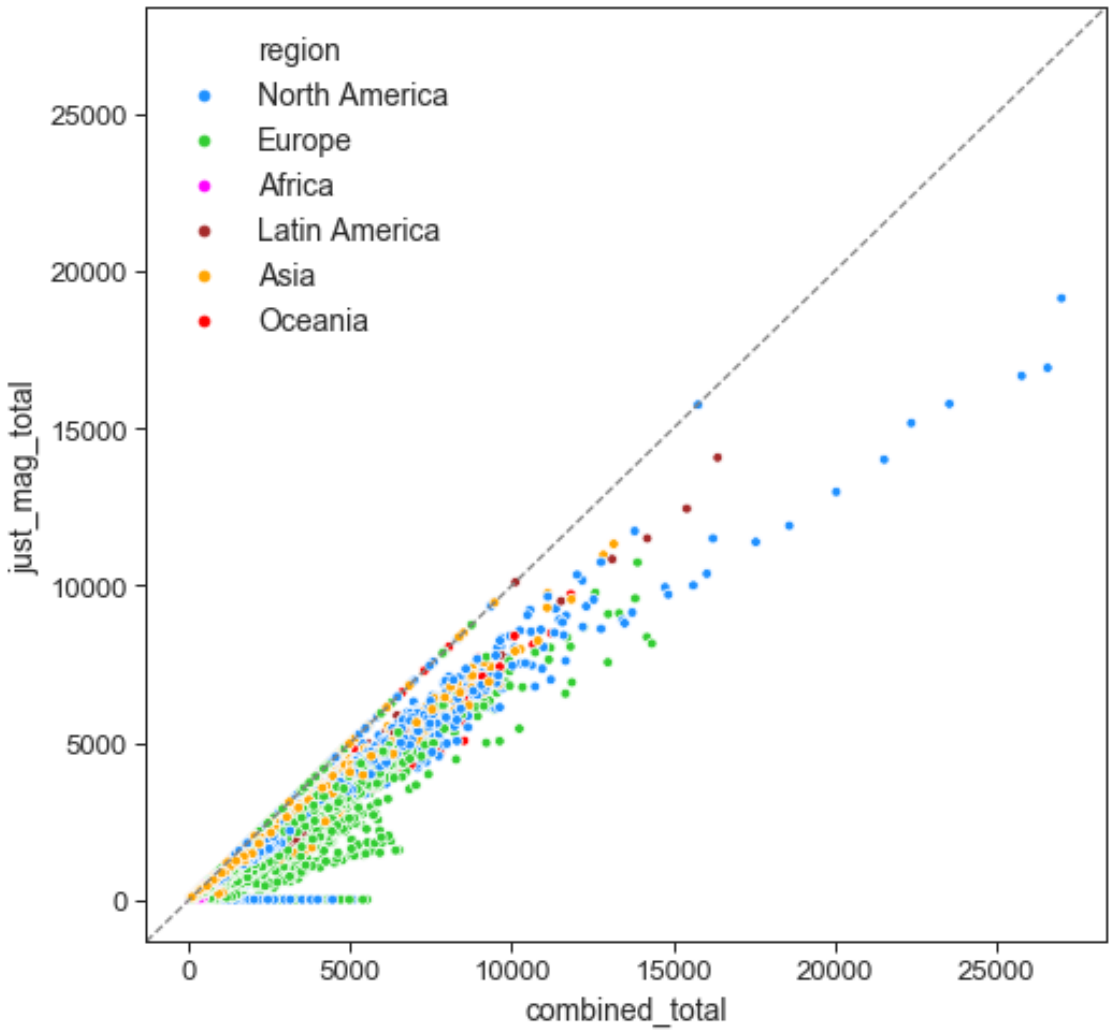
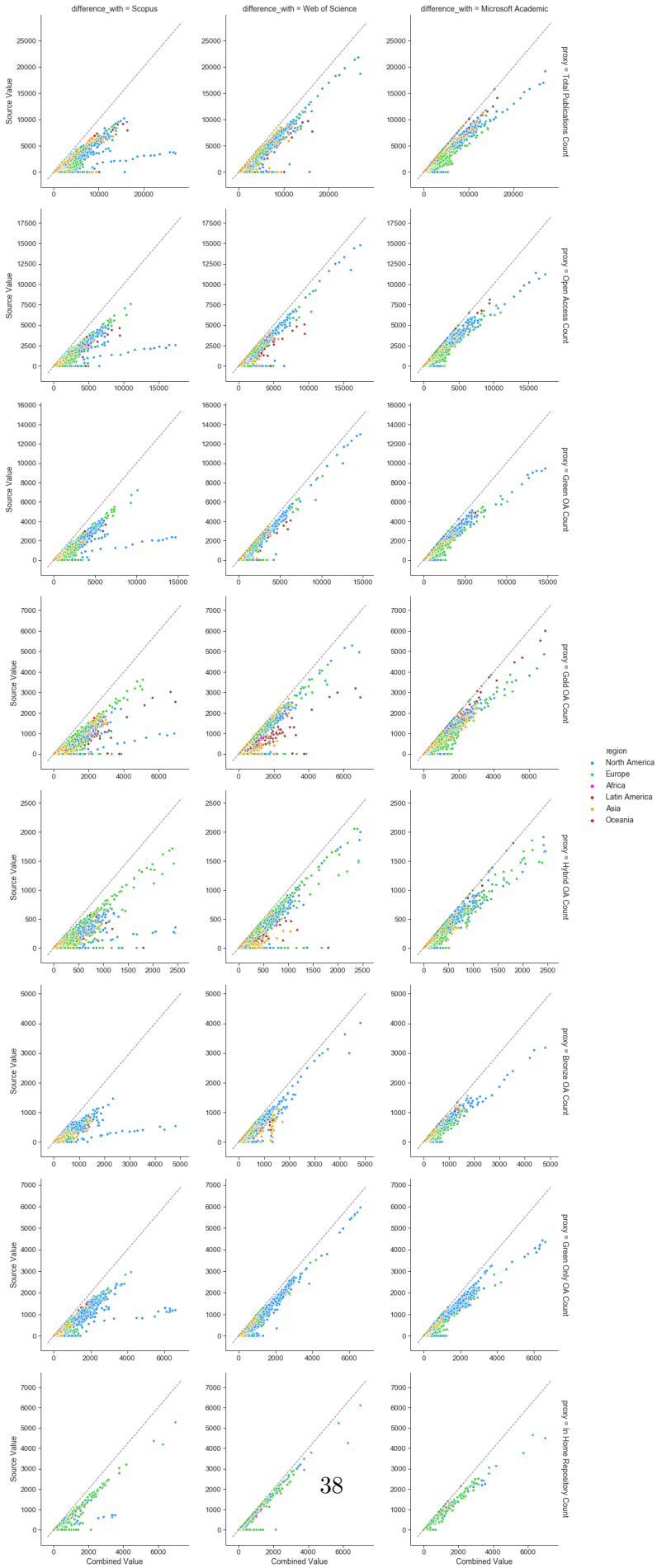


Figure 18 compares the counts in total publication and various open access categories obtained throuh the combined datset against those derived from only individual sources. We first observe that alomst all dots lie below the diagonal line (few sit on the line), indicating that the combined dataset gives a significantly high number of counts in all categories as we would expect. For gold open access counts, we observe that Latin Amrican universities lie much further away from the diagonal line in the comparison against Scopus and Web of Science. This is an indication of the effect of the inclusion of Microsoft Academic have on universities from this region. It is also interesting to note that the number of outputs in the green open access and green in home reposities are dominated by European and North American universities.

Figure 18: Scatterplots of the combined dataset against Web of Science, Scopus and Microsoft Academic, in terms of publication counts (in total publication, total OA, gold OA, green OA, hybrid OA, bronze OA, green only OA, and green in home repo OA) for all years in the dataframe.


difference_with = Scopus
difference_with = Web of Science
difference_with = Microsoft Academic
proxy = Total Publications Count
proxy = Open Access Count
proxy = Green OA Count
proxy = Gold OA Count
proxy = Hybrid OA Count
proxy = Bronze OA Count
proxy = Green Only OA Count
proxy = In Home Repository Count
Source Value
Combined Value
region
North America
Europe
Africa
Latin America
Asia
Oceania

In our data workflow, Scopus consistently give Harvard a lower number of outputs. This is shown by the slowly increasing line (of blue dots) on the bottom right of the plots in the first column above. This is due to the affiliation identifier in Scopus, where Harvard has multiple identifiers. The identifier in our database points to "Harvard University" but that does not cover all of "Harvard Medical School" which has a higher number of total outputs. This emphasises the importance of integrating multiple data sources to address differences between data sources.

Figure 19 presents corresponding scatterplots in terms of the difference categories of open access percentages. Several interesting patterns arise that highlights regional differences across the data sources. For the total open access percentages, we see almost a clear divide between European and North American universities, and the rest of the world (across the diagonal line), in the comparison of the combined dataset against Scopus and Web of Science. This is a clear evidence for the bias driven by Web of Science and Scopus. This is further explained in detail by observing the detailed open access categories. For example, the gold open access percentages for Latin American and Asian universities are significantly higher when using the combined dataset, in contrast to using only Web of Science or Scopus. Similarly, Asian universities are assigned significantly higher number of bronze open access through the combined dataset.

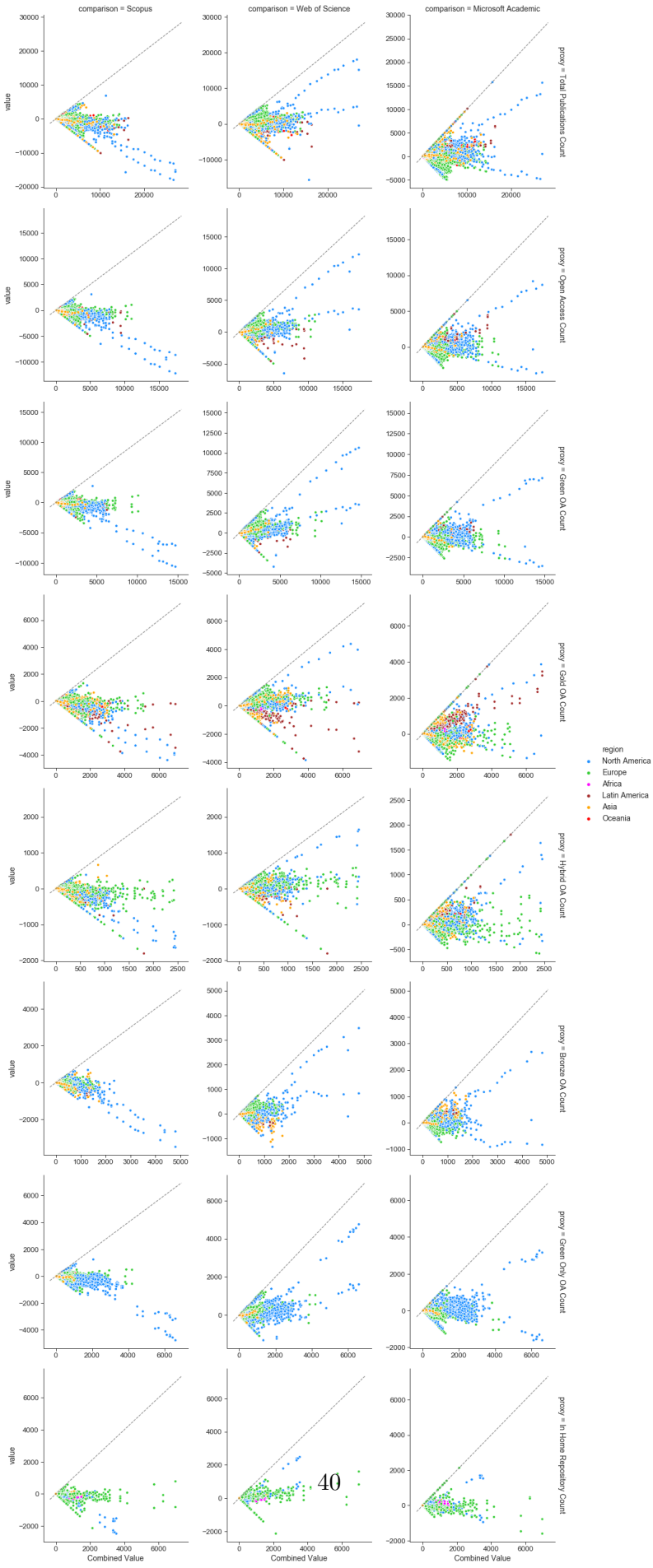
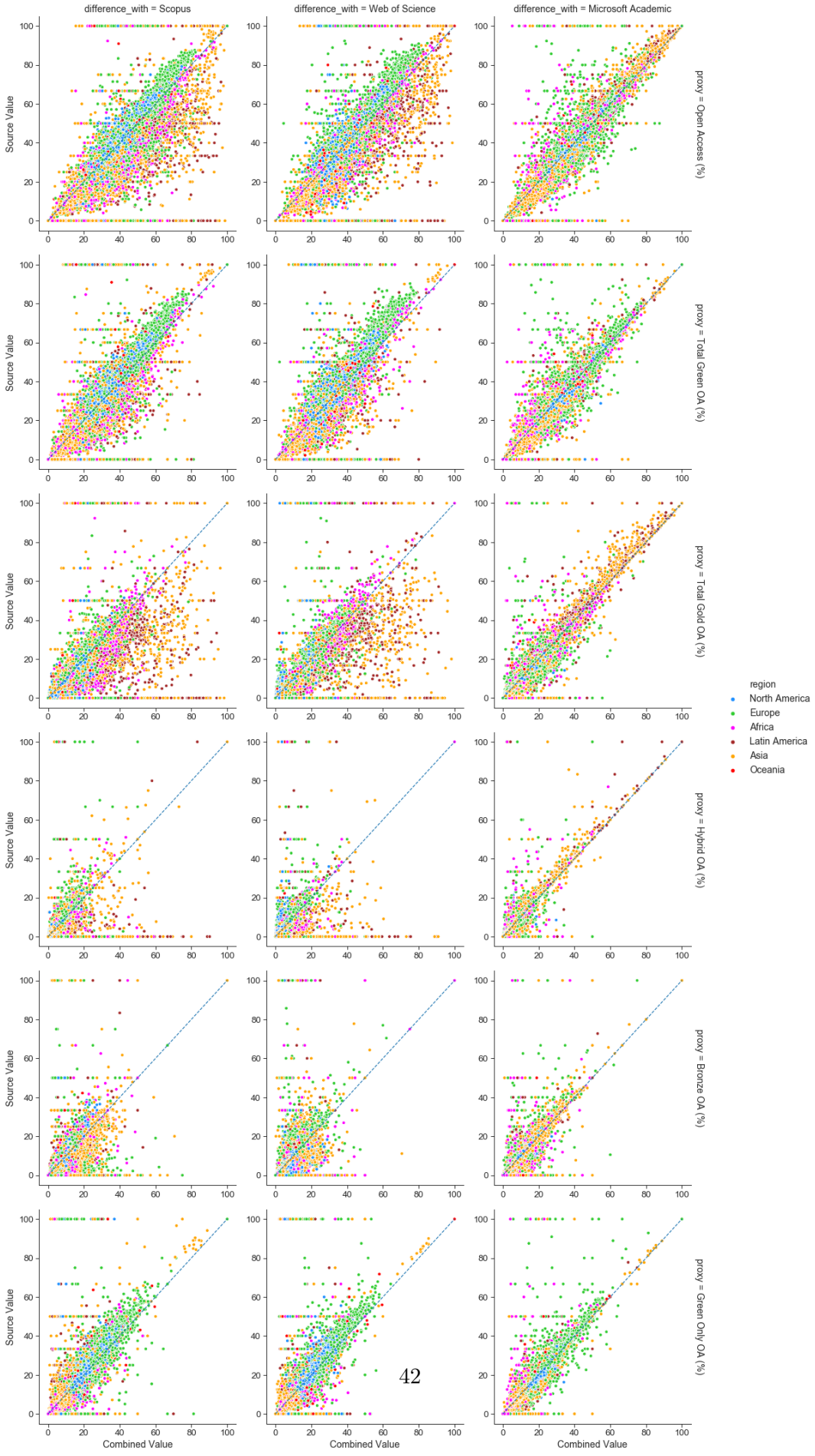

comparison = Scopus
comparison = Web of Science
comparison = Microsoft Academic
proxy = Total Publications Count
proxy = Open Access Count
proxy = Green OA Count
proxy = Gold OA Count
proxy = Hybrid OA Count
proxy = Bronze OA Count
proxy = Green Only OA Count
proxy = In Home Repository Count
value
Combined Value
region
North America
Europe
Africa
Latin America
Asia
Oceania

Figure 19: Scatterplots of the combined dataset against Web of Science, Scopus and Microsoft Academic, in terms of OA percentages (in total OA, gold OA, green OA, hybrid OA, bronze OA, and green only OA) for all years in the dataframe.


difference_with = Scopus
difference_with = Web of Science
difference_with = Microsoft Academic
proxy = Open Access (%)
proxy = Total Green OA (%)
proxy = Total Gold OA (%)
proxy = Hybrid OA (%)
proxy = Bronze OA (%)
proxy = Green Only OA (%)
Source Value
Combined Value
region
North America
Europe
Africa
Latin America
Asia
Oceania

3 Sensitivity analysis on the use of different Unpaywall snapshot versions

In this section, we proceed to analyse the levels of sensitivity asscociated with the use of different Unpaywall data dumps. For these analyses, we maintain the use of the combined dataset, but use Unpaywall data dumps from different dates to determine open access status.

Figures 20, 21 and 22, compares our latest Unpaywall data dump against earlier versions for calculating total open access percentages, gold open access percentage, and green open access percentages, respectively, for 2017. Figure 20 shows that there are clearly more differences as we move towards earlier versions of Unpaywall, with many universities moving further away from the diagonal. However, Figures 21 and 22 further highlight that these differences are significantly due to the changes in green open access levels. This is an indication of Unpaywalll's improving ability to capture data recorded by research repositories.

[38]:

	id	country	region	subregion	published_year	country_code	Later Value	Earlier Value	Difference	Later Release
0	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2005	USA	2297.000000	2297.000000	0.000000	unpaywall_2019_11_22
1	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2017	USA	5244.000000	5243.000000	1.000000	unpaywall_2019_11_22
2	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2019	USA				
3	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2010	USA				
4	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	North America	Northern America	2018	USA				
…	…	…	…	…	…	…				
22514	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	Asia	South-eastern Asia	2015	MYS				
22515	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	Asia	South-eastern Asia	2016	MYS				
22516	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	Asia	South-eastern Asia	2017	MYS				
22517	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	Asia	South-eastern Asia	2018	MYS				
22518	ffe0002dce72a38ae58852fe3e6f1b57	Malaysia	Asia	South-eastern Asia	2019	MYS				

```
2      3060.000000    2633.000000  427.000000  unpaywall_2019_11_22
3      3897.000000    3897.000000    0.000000  unpaywall_2019_11_22
4      5549.000000    5548.000000    1.000000  unpaywall_2019_11_22
…              …              …           …                     …
22514     2.923149       2.963311   -0.040162  unpaywall_2018_09_24
22515     2.066698       2.578528   -0.511830  unpaywall_2018_09_24
22516     2.078138       2.405641   -0.327503  unpaywall_2018_09_24
22517     1.092299       2.355372   -1.263073  unpaywall_2018_09_24
22518     1.923077       0.544959    1.378118  unpaywall_2018_09_24

            Earlier Release                    proxy
0      unpaywall_2019_08_16  Total Publications Count
1      unpaywall_2019_08_16  Total Publications Count
2      unpaywall_2019_08_16  Total Publications Count
3      unpaywall_2019_08_16  Total Publications Count
4      unpaywall_2019_08_16  Total Publications Count
…                       …                  …
22514  unpaywall_2019_02_21         Green Only OA (%)
22515  unpaywall_2019_02_21         Green Only OA (%)
22516  unpaywall_2019_02_21         Green Only OA (%)
22517  unpaywall_2019_02_21         Green Only OA (%)
22518  unpaywall_2019_02_21         Green Only OA (%)

[6306664 rows x 12 columns]
```

Figure 20: Scatterplots contrasting total OA% between various versions of Unpaywall data dumps.

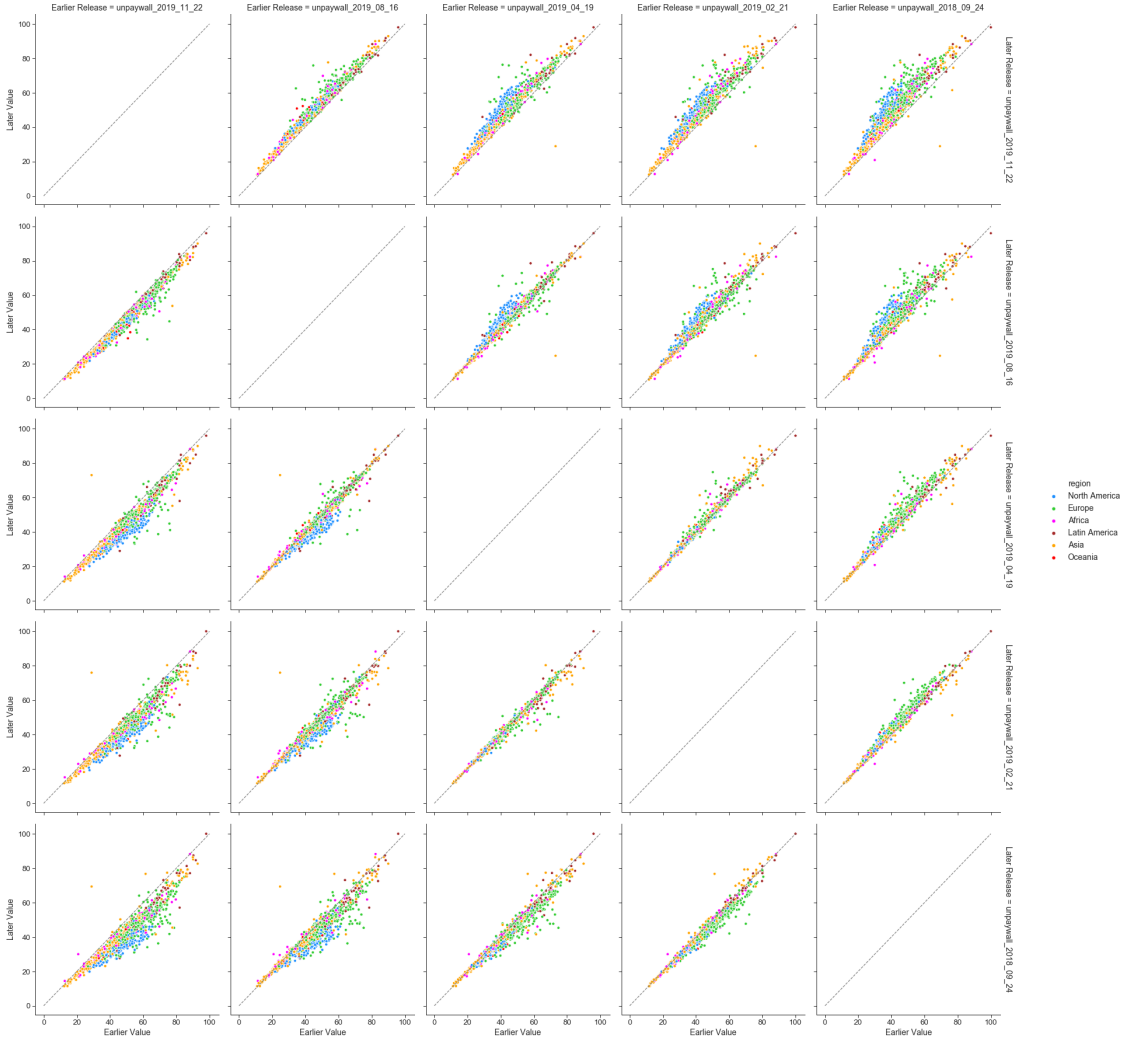


Figure 21: Scatterplots contrasting gold OA% between various versions of Unpaywall data dumps.

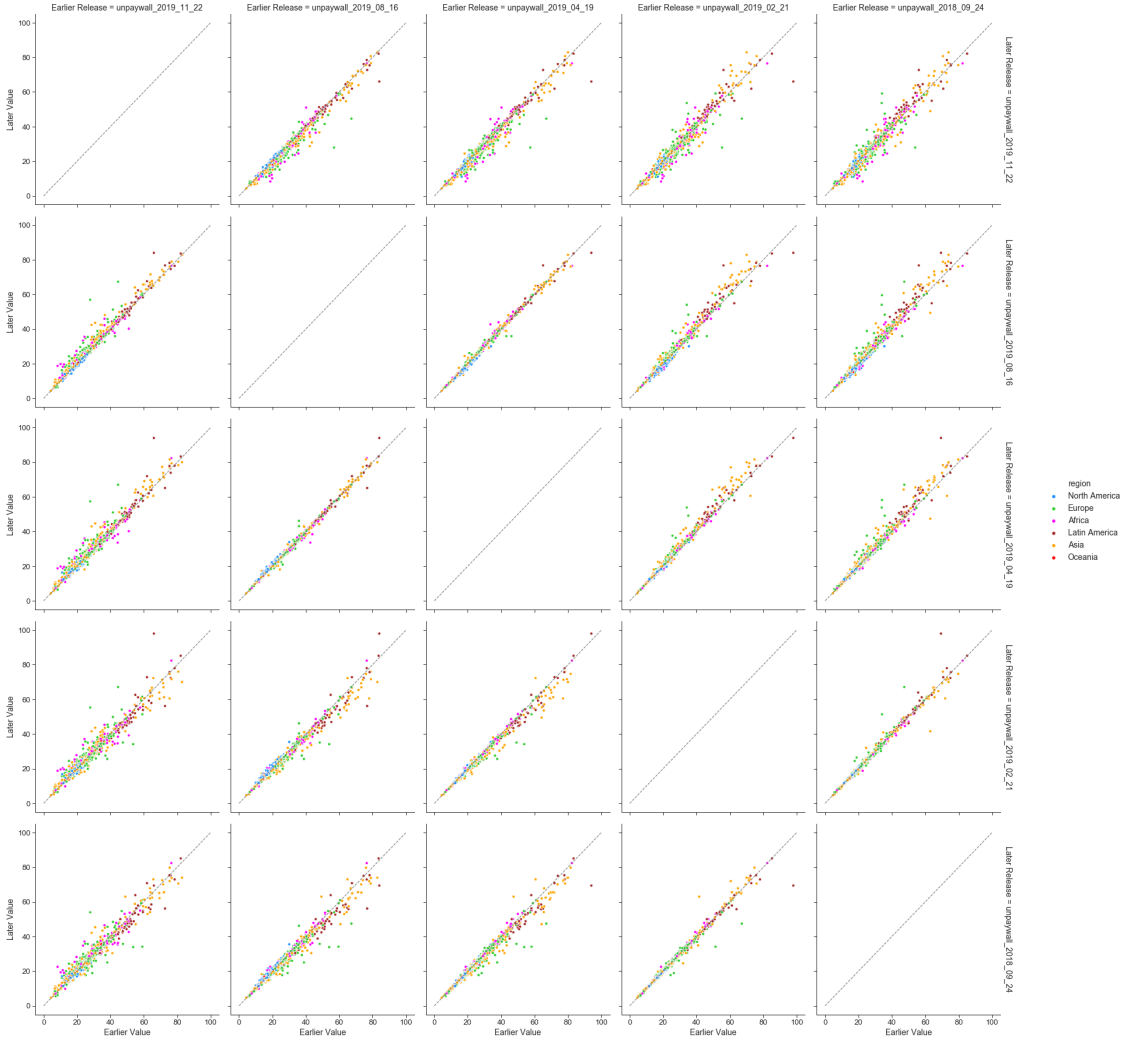
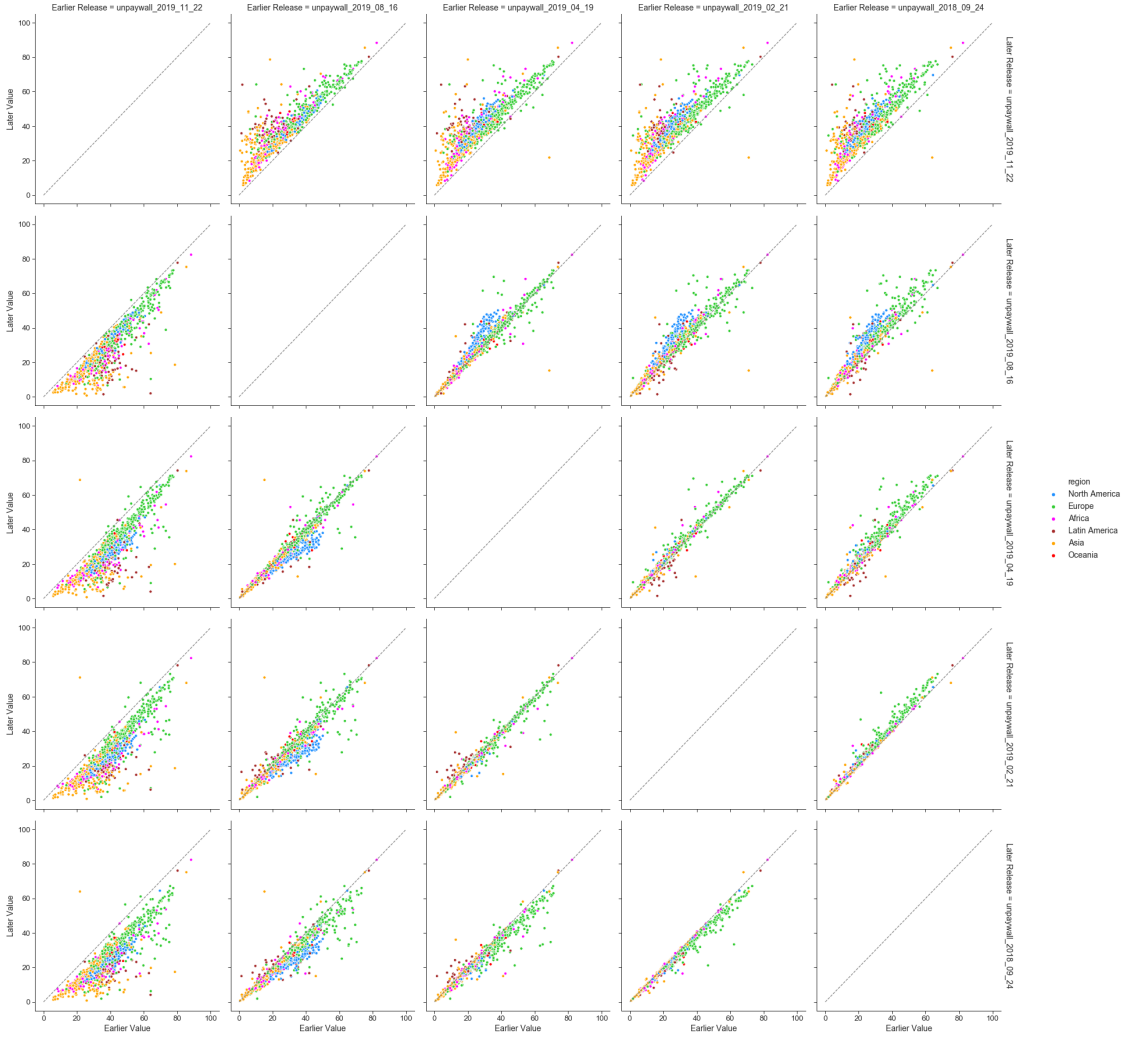


Figure 22: Scatterplots contrasting green OA% between various versions of Unpaywall data dumps.



Figures 23 to 27 depict the differences in total, gold, green, hybrid and bronze open access percentages (respectively) between the latest version of Unpaywall and several earlier versions. As observed earlier, larger differences are observed for total open access and green open access, as opposed to other open access categories.

Two other interesting phenomenon are evidenced. Firstly, Unpaywall seems to be capturing a backfilling of historical data in particularly through the green and bronze routes. Secondly, we seem to observe the effects of embargos on self-archiving through the sudden jumps in 2018 for green open access, as we move towards earlier data dumps. To a lesser degree, some evidence of publisher embargos are shown in the gold open access.

Figure 23: Comparing the most recent version of Unpaywall against earlier versions in terms of total OA%, for all years.

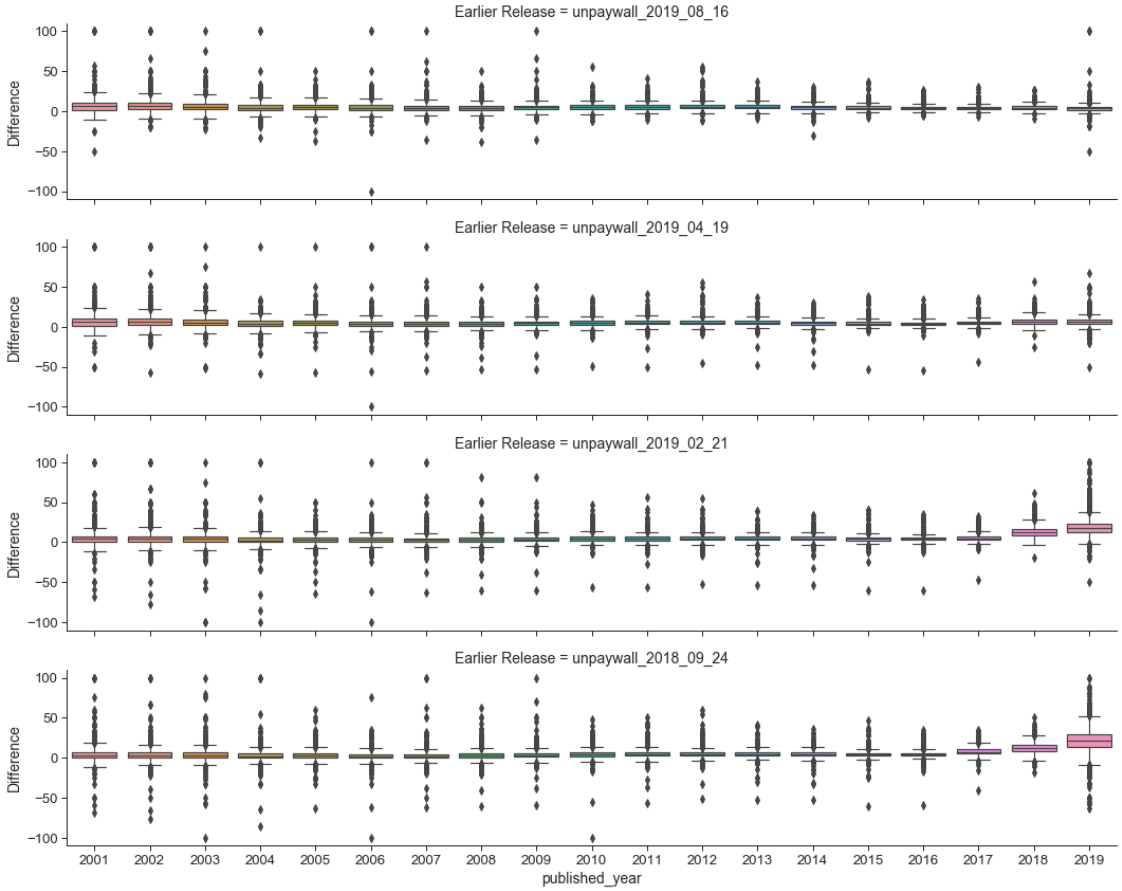


Figure 24: Comparing the most recent version of Unpaywall against earlier versions in terms of gold OA%, for all years.

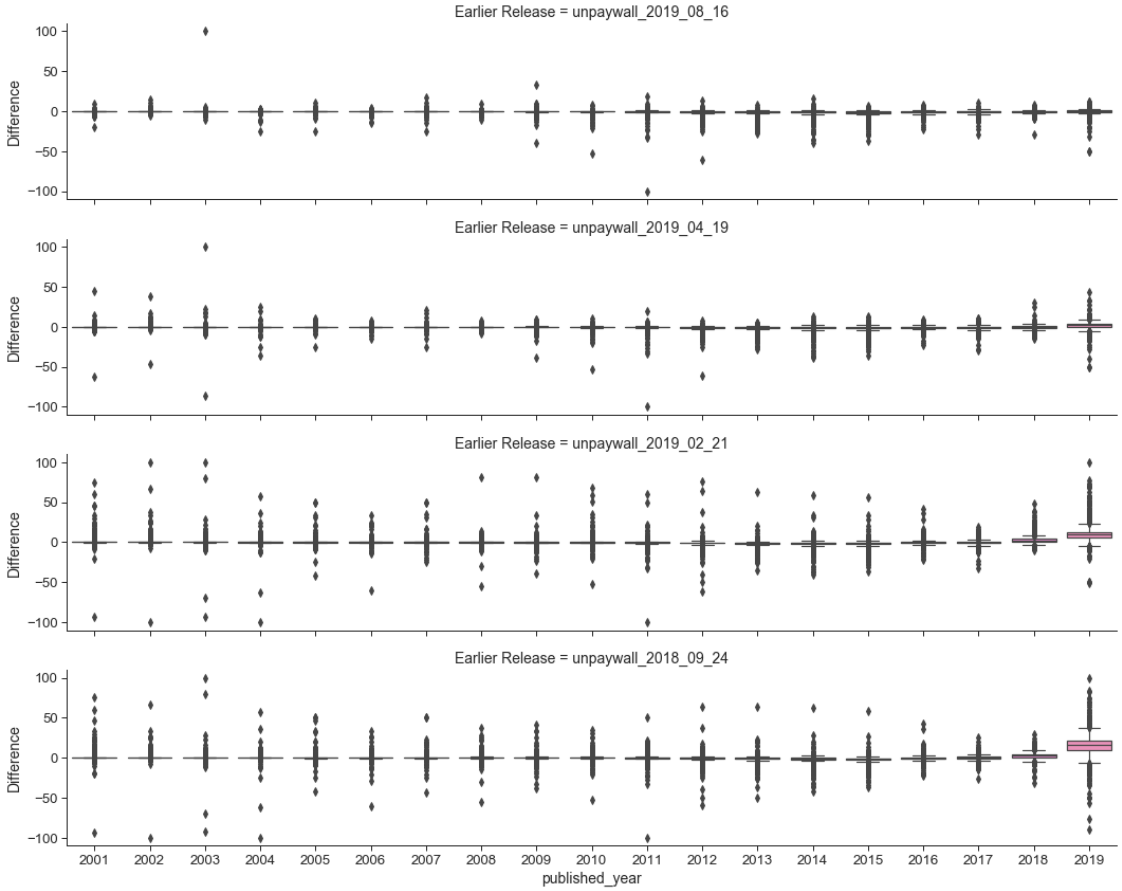


Figure 25: Comparing the most recent version of Unpaywall against earlier versions in terms of green OA%, for all years.

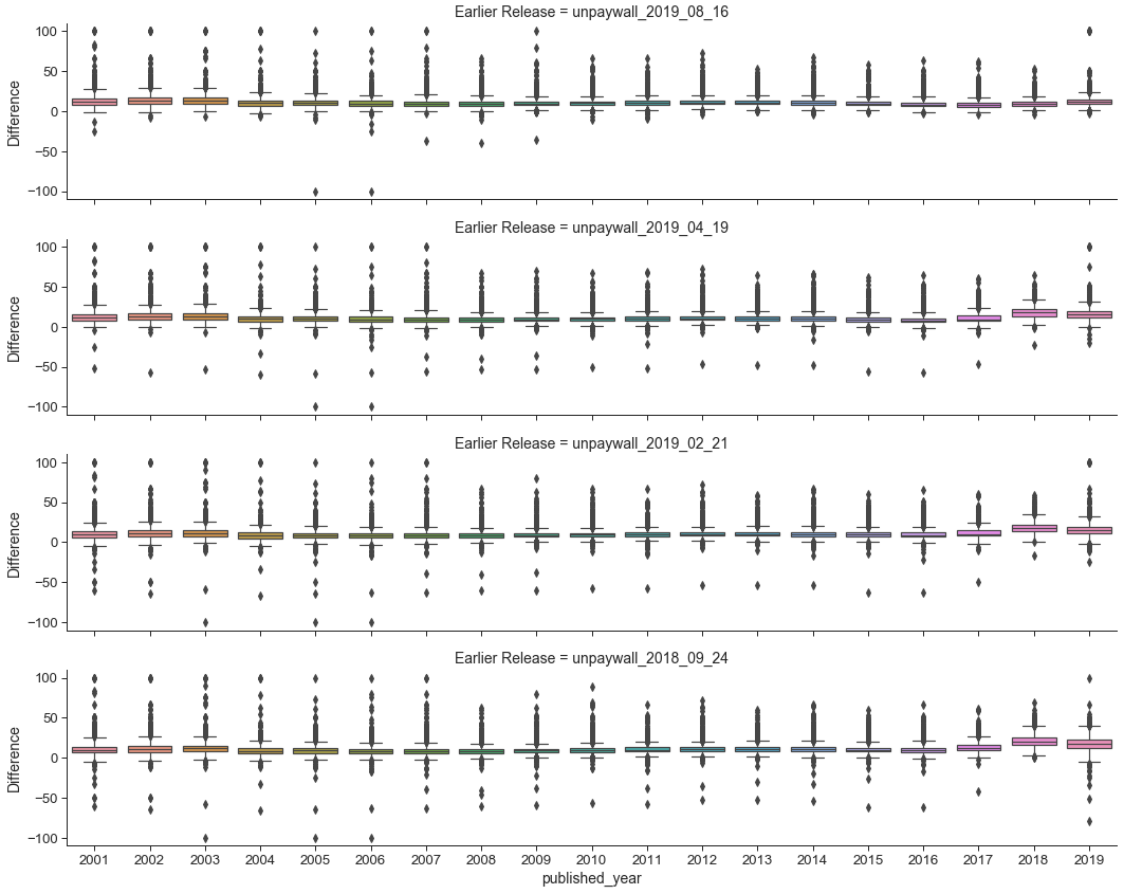


Figure 26: Comparing the most recent version of Unpaywall against earlier versions in terms of hybrid OA%, for all years.

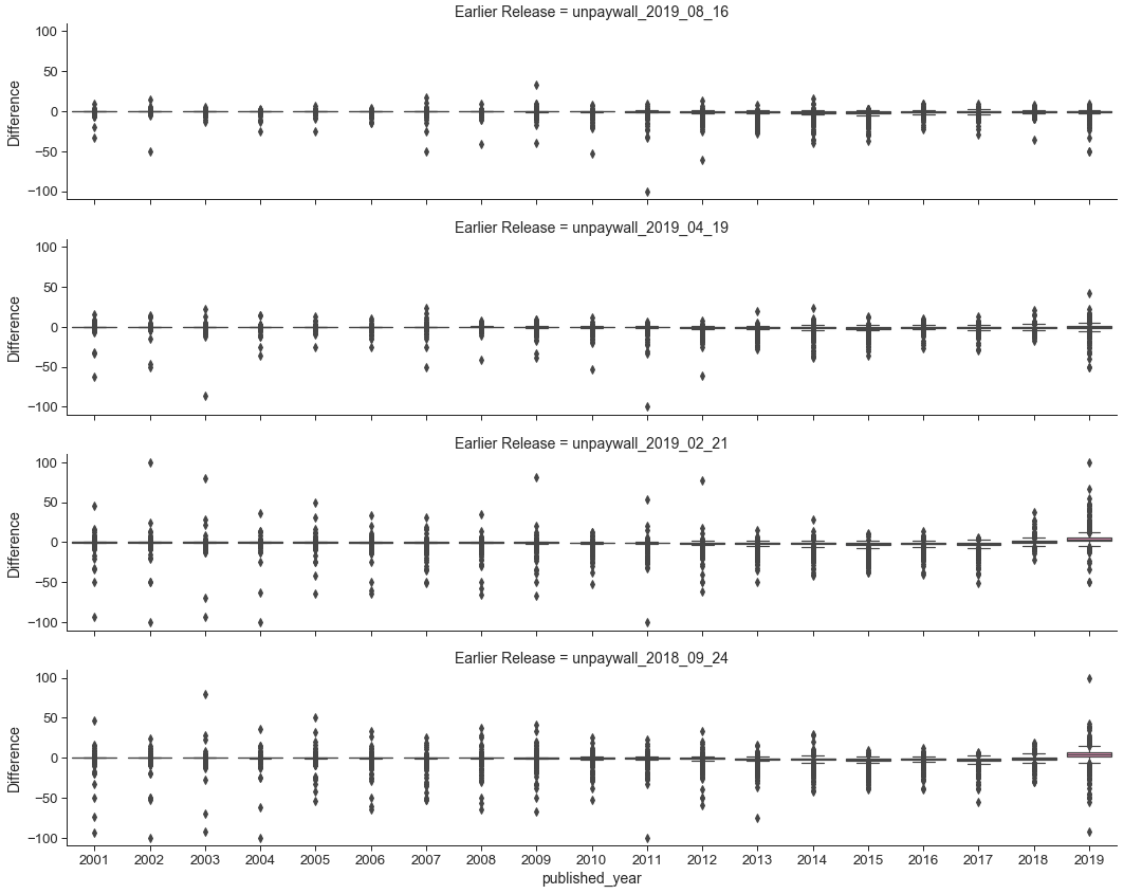
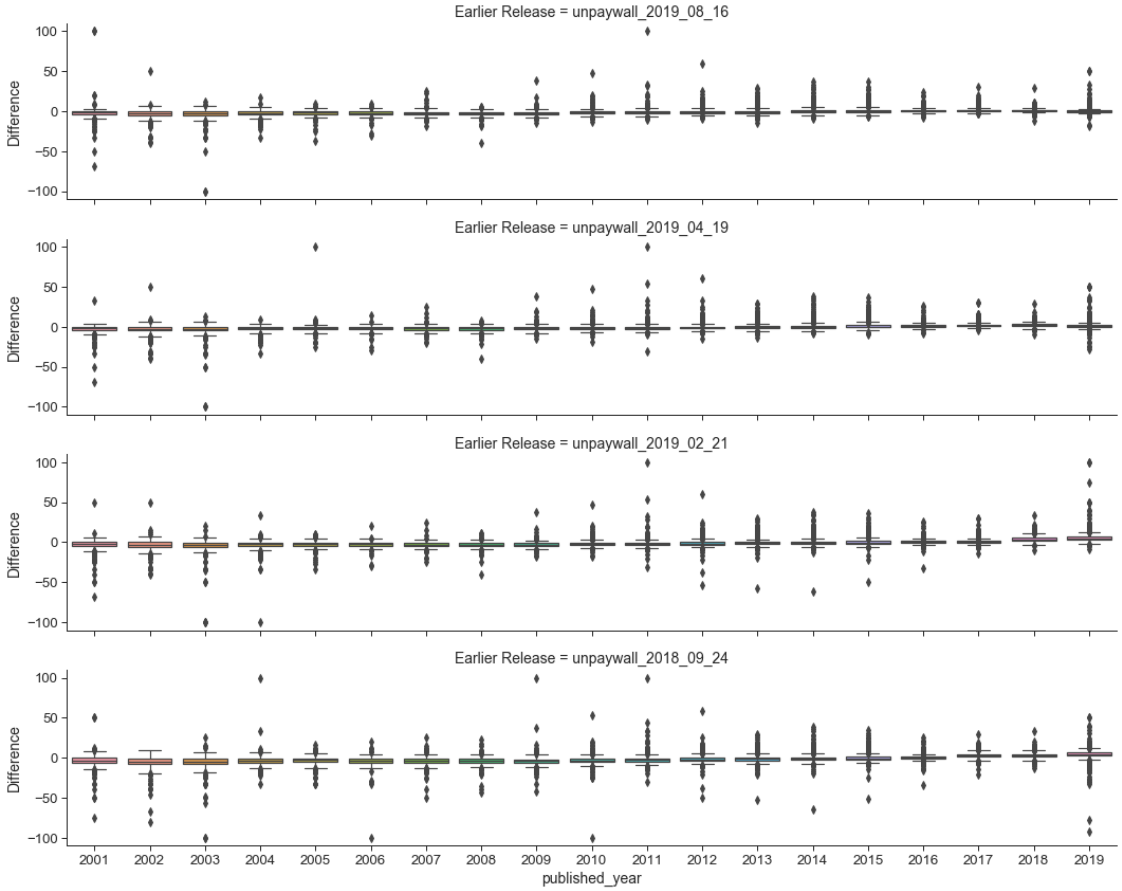
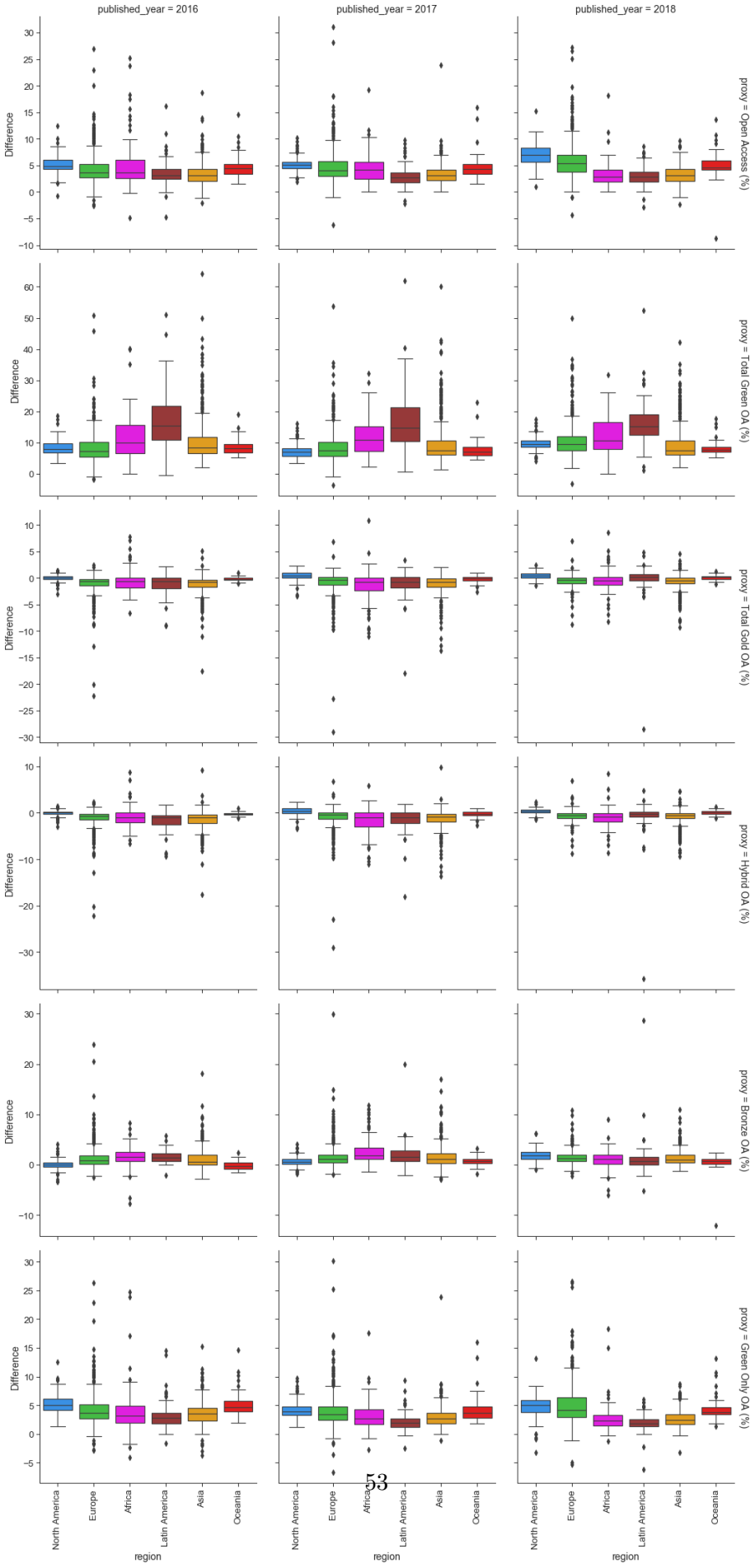
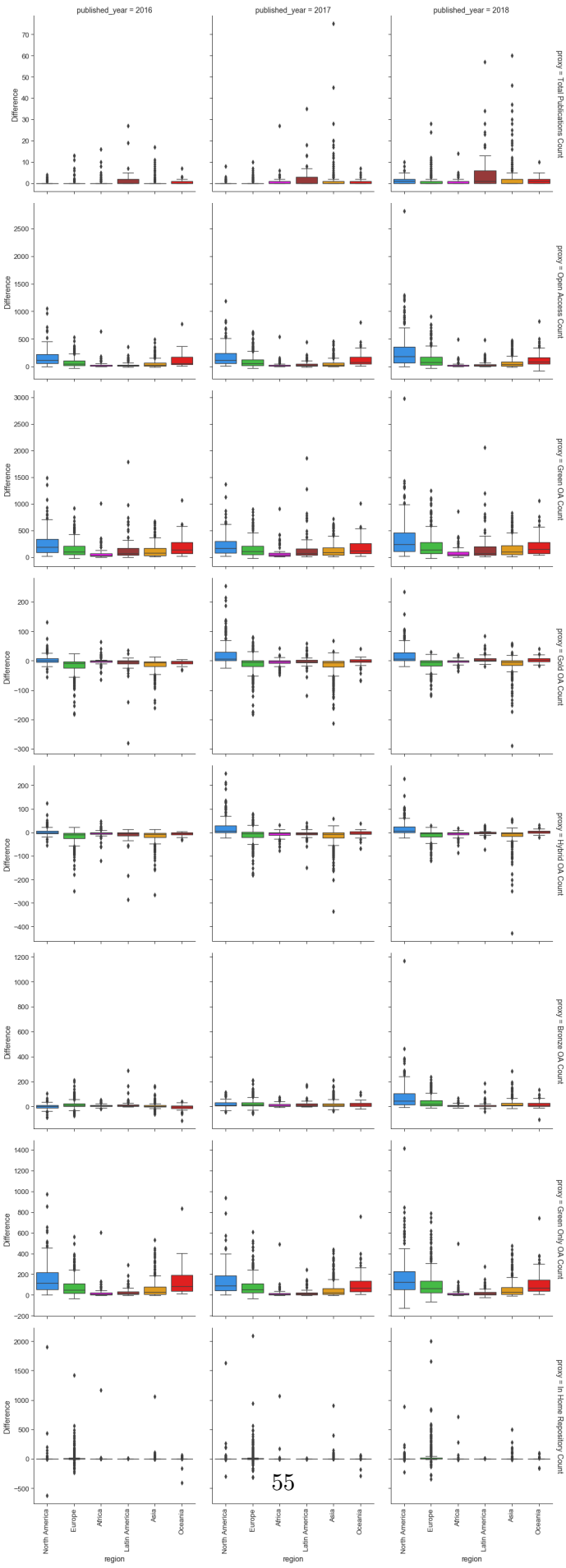
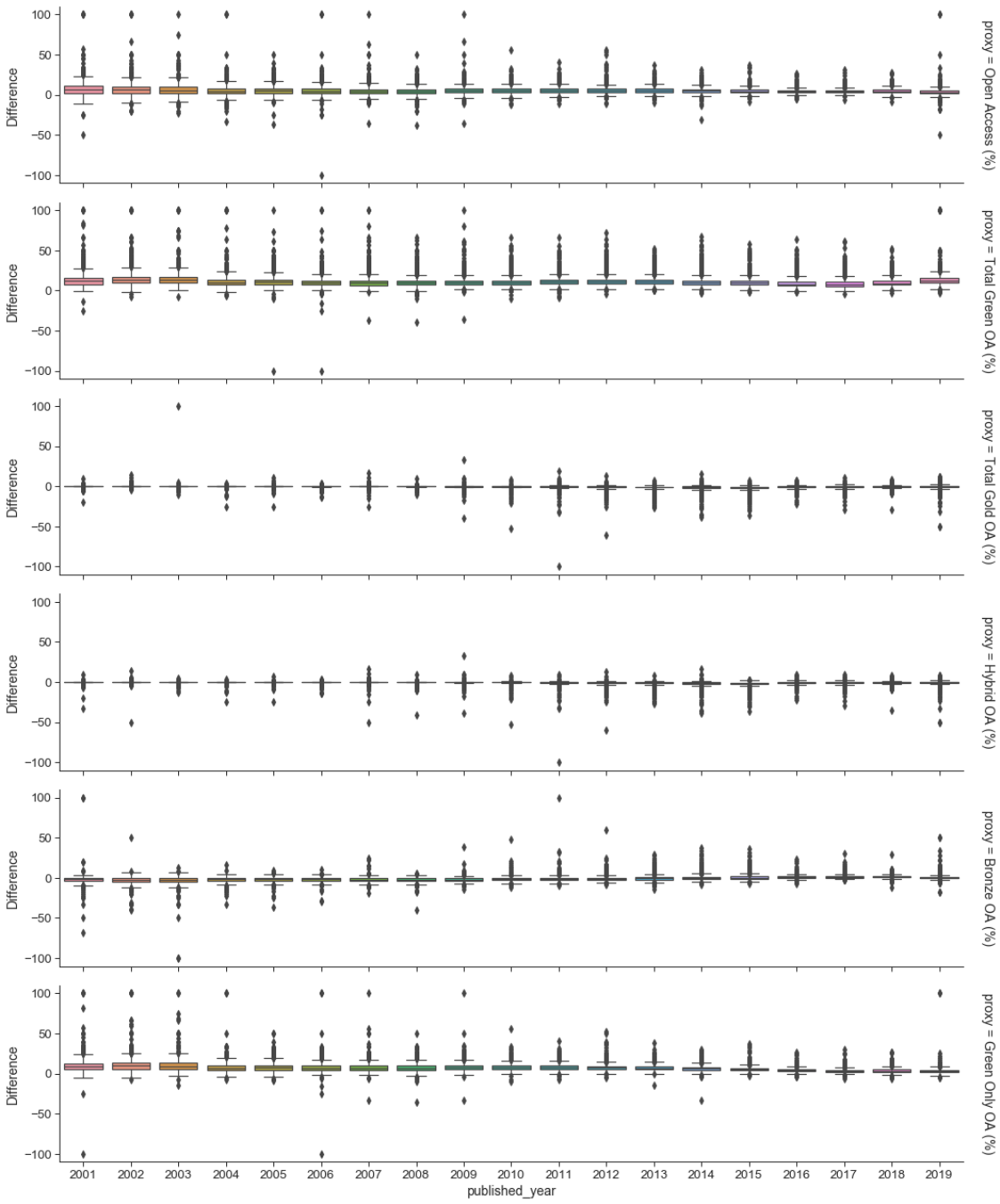


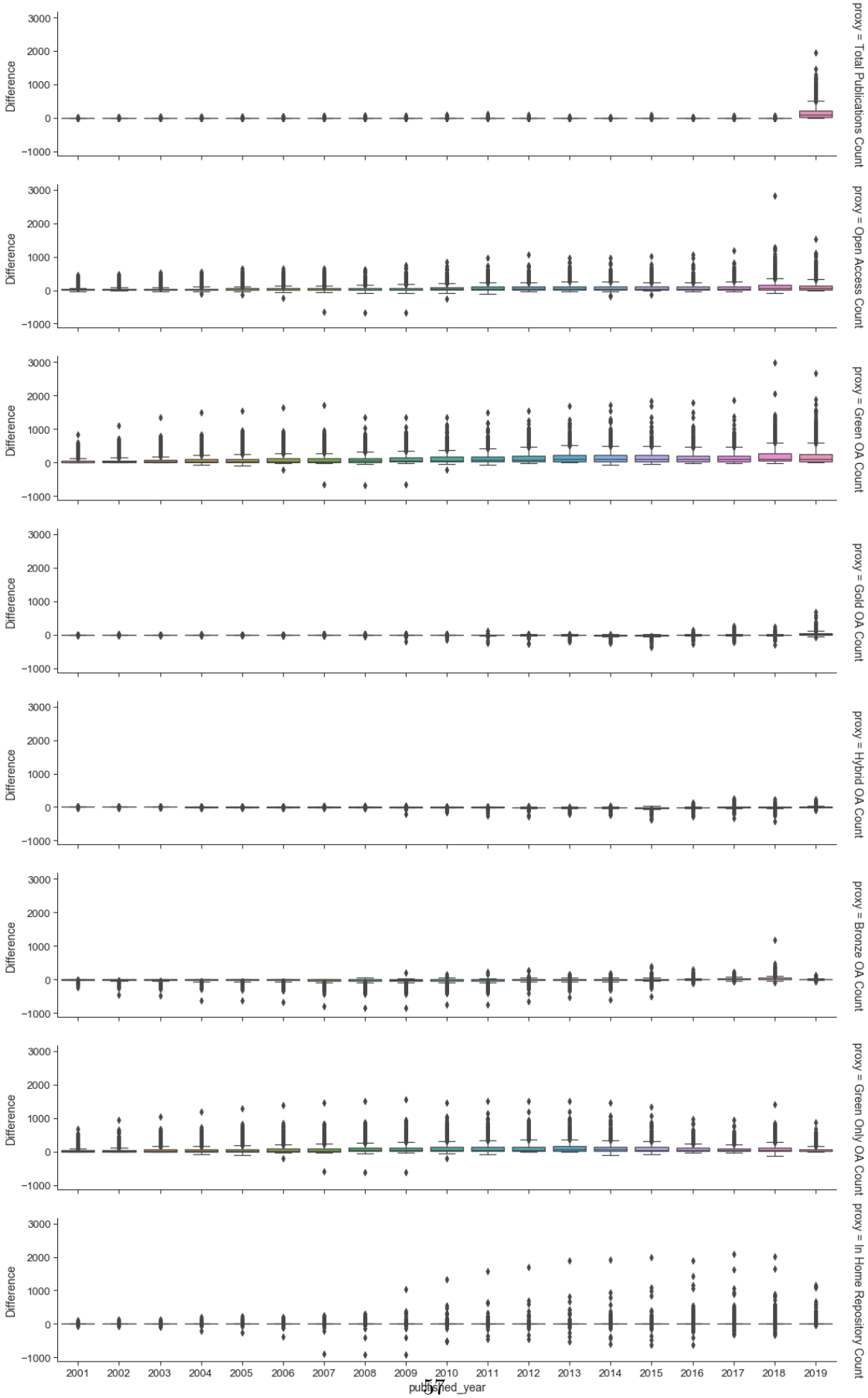
Figure 27: Comparing the most recent version of Unpaywall against earlier versions in terms of bronze OA%, for all years.




published_year = 2016
published_year = 2017
published_year = 2018
proxy = Open Access (%)
proxy = Total Green OA (%)
proxy = Total Gold OA (%)
proxy = Hybrid OA (%)
proxy = Bronze OA (%)
proxy = Green Only OA (%)
Difference
region
North America
Europe
Africa
Latin America
Asia
Oceania


published_year = 2016
published_year = 2017
published_year = 2018
proxy = Total Publications Count
proxy = Open Access Count
proxy = Green OA Count
proxy = Gold OA Count
proxy = Hybrid OA Count
proxy = Bronze OA Count
proxy = Green Only OA Count
proxy = In Home Repository Count
Difference
region
North America
Europe
Africa
Latin America
Asia
Oceania


proxy = Open Access (%)
proxy = Total Green OA (%)
proxy = Total Gold OA (%)
proxy = Hybrid OA (%)
proxy = Bronze OA (%)
proxy = Green Only OA (%)
Difference
100
50
0
−50
−100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
published_year


proxy = Total Publications Count
proxy = Open Access Count
proxy = Green OA Count
proxy = Gold OA Count
proxy = Hybrid OA Count
proxy = Bronze OA Count
proxy = Green Only OA Count
proxy = In Home Repository Count
Difference
3000
2000
1000
0
−1000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
published_year

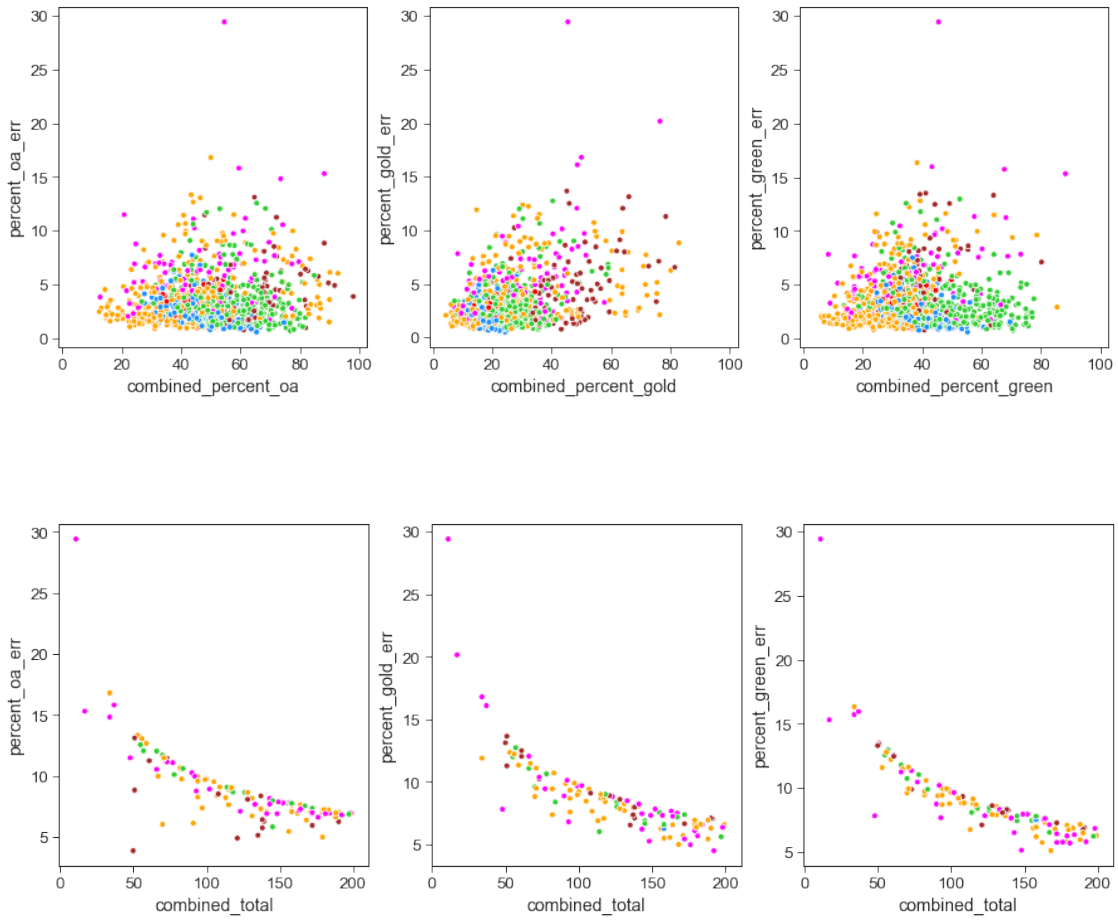
4 Sensitivity analysis on sample size, margin of error and confidence levels

In this section we focus on analysing the sample size, margin of error and confidence levels associated with each university's open access percentage estimates. In the corresponding main article (Huang et al, 2020b), margins of error were used to exclude universities that we have very little confidence for in terms of their open access levels produced through the data workflow. That means universities with a margin of error below a certain cut-off were not included in the list of Top 100 and some of the subsequent analysis. This is in addition to excluding universities that do not satify the conventional conditions for normal approximation to proportions (i.e., $np > 5$ and $n(1-p) > 5$, where n is the sample size and p is the sample proportion for success).

For the choice of a cut-off level, we analysed the margins of error (equivalently, half of the length of the corresponding confidence intervals) at a number of confidence levels (i.e., 95%, 99% and 99.5%). In should be noted that we used the Šidák correction to control for the familywise error rate in multiple comparisons. This means that the multiple 95% confidence intervals calculated across the Top 100 universities are essentially 99.95% individual confidence intervals for each university. Hence, it justifies comparing the associated margins of error across a number of confidence levels in this range.

In Figures 28, 29 and 30, we plot the margins of error (at 95%, 99% and 99.95% respectively) against total, gold and green open access percentages, and against the total number of output, for each university. The aim is the spot a cut-off point for the margin of error when the data starts to behave more extremely. Subjectively, we consider the cut-off levels at 9.8, 12.8 and 17 for the three difference confidence levels. These correspond to approximately a sample size cut-off at 100 outputs, with a few exceptions. We consider this as a good starting point as we are focussed on research-intensive universities (Huang et al., 2020b) and we also would like a certain level of confidence about the open access estimates we provide. It is our aim to take a deeper exploration on issues around multiple comparisons (e.g., more advanced methods to cater for multiple comparisons) and principles for inclusion and exclusion.

Figure 28: 95% Margins of error versus open access percentages and the total number of publications for 2017.



[52]:

	id	country	\
1	4ae1dfc3e9dbfa802d5aa94ecaacd8f3	United States of America	
24	3f3dac00919266d9156e20abf3e05d0a	Netherlands	
44	293018e797a8353364a328b1af4518b5	United States of America	
70	4cc38006c47b39f0e8dbfa054ea44943	Ghana	
78	0fd955e857845e0179b66816ae907e17	Argentina	
…	…	…	
22441	9aaa25d74d1694a0df5a1e3ad79d41bf	Germany	
22455	5a608e2f933f8c5e9aa494d5ee000607	United States of America	
22479	3d18990445717f490a7037ce0c43b223	United States of America	
22492	92f84dd0c88bb4c0823383f8026e0b37	South Korea	
22506	8fac4241d860e90d4eced37eae996fd5	China	

	country_code	region	subregion	\
1	USA	North America	Northern America	
24	NLD	Europe	Western Europe	

```
44             USA  North America                  Northern America
70             GHA         Africa                Sub-Saharan Africa
78             ARG  Latin America  Latin America and the Caribbean
…              …              …                               …
22441          DEU         Europe                   Western Europe
22455          USA  North America                  Northern America
22479          USA  North America                  Northern America
22492          KOR           Asia                      Eastern Asia
22506          CHN           Asia                      Eastern Asia

       published_year  combined_total  combined_oa  combined_green  \
1                2017            5245         3249            2755
24               2017            7283         4800            3992
44               2017            1238          481             403
70               2017             215          124             100
78               2017             487          185             158
…                …               …            …               …
22441            2017             654          268             222
22455            2017            2855         1218            1012
22479            2017            2191          906             756
22492            2017            1707          804             646
22506            2017           10808         4528            3642

       combined_gold  …  just_mag_percent_oa  just_mag_percent_green  \
1               1159  …            61.238464               51.741590
24              2783  …            65.999629               54.622939
44               184  …            40.062112               33.022774
70                88  …            56.818182               44.886364
78               113  …            35.849057               29.380054
…                …    …              …                     …
22441            138  …            29.539295               22.493225
22455            538  …            38.307350               30.567929
22479            371  …            38.609272               32.450331
22492            520  …            39.951768               30.225080
22506           2856  …            34.946041               27.319025

       just_mag_percent_gold  just_mag_percent_hybrid  \
1                  21.911283                 7.948794
24                 39.281082                20.307578
44                 15.320911                 4.658385
70                 38.068182                 6.818182
78                 22.371968                 4.312668
…                      …                    …
22441              11.382114                 3.523035
22455              15.534521                 4.732739
22479              15.430464                 4.304636
22492              26.527331                 8.279743
```

```
22506              19.679884                  4.110586

       just_mag_percent_bronze  just_mag_percent_green_only  \
1                    14.796070                    24.531110
24                   11.784325                    14.934223
44                    7.867495                    16.873706
70                   10.227273                     8.522727
78                    3.773585                     9.703504
…                        …                            …
22441                 6.233062                    11.924119
22455                 8.129176                    14.643653
22479                 9.006623                    14.172185
22492                 6.189711                     7.234727
22506                 7.396629                     7.869528

       just_mag_percent_in_home_repo  percent_oa_err  percent_gold_err  \
1                           0.057197        1.313993          1.122868
24                          0.000000        1.088678          1.115972
44                          0.807754        2.715161          1.981547
70                          0.000000        6.604348          6.572666
78                          0.000000        4.310741          3.749192
…                              …              …                 …
22441                       0.000000        3.769211          3.127187
22455                       0.000000        1.814241          1.434502
22479                       0.000000        2.062093          1.570418
22492                       0.000000        2.367978          2.183398
22506                       0.000000        0.930189          0.831295

       percent_green_err
1               1.351445
24              1.143009
44              2.610182
70              6.667259
78              4.158044
…                  …
22441           3.629174
22455           1.754688
22479           1.990579
22492           2.300805
22506           0.891141

[1206 rows x 69 columns]
```

Figure 29: 99% Margins of error versus open access percentages and the total number of publications for 2017.

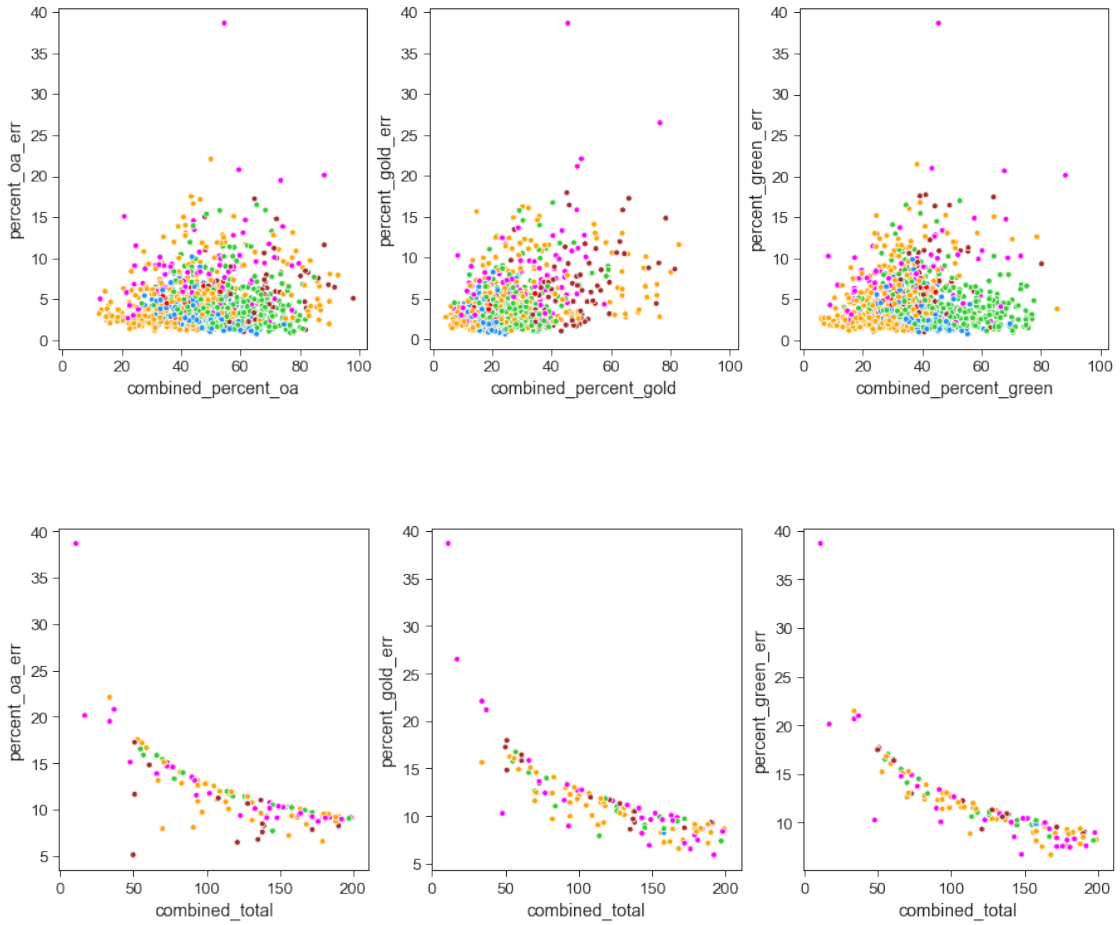
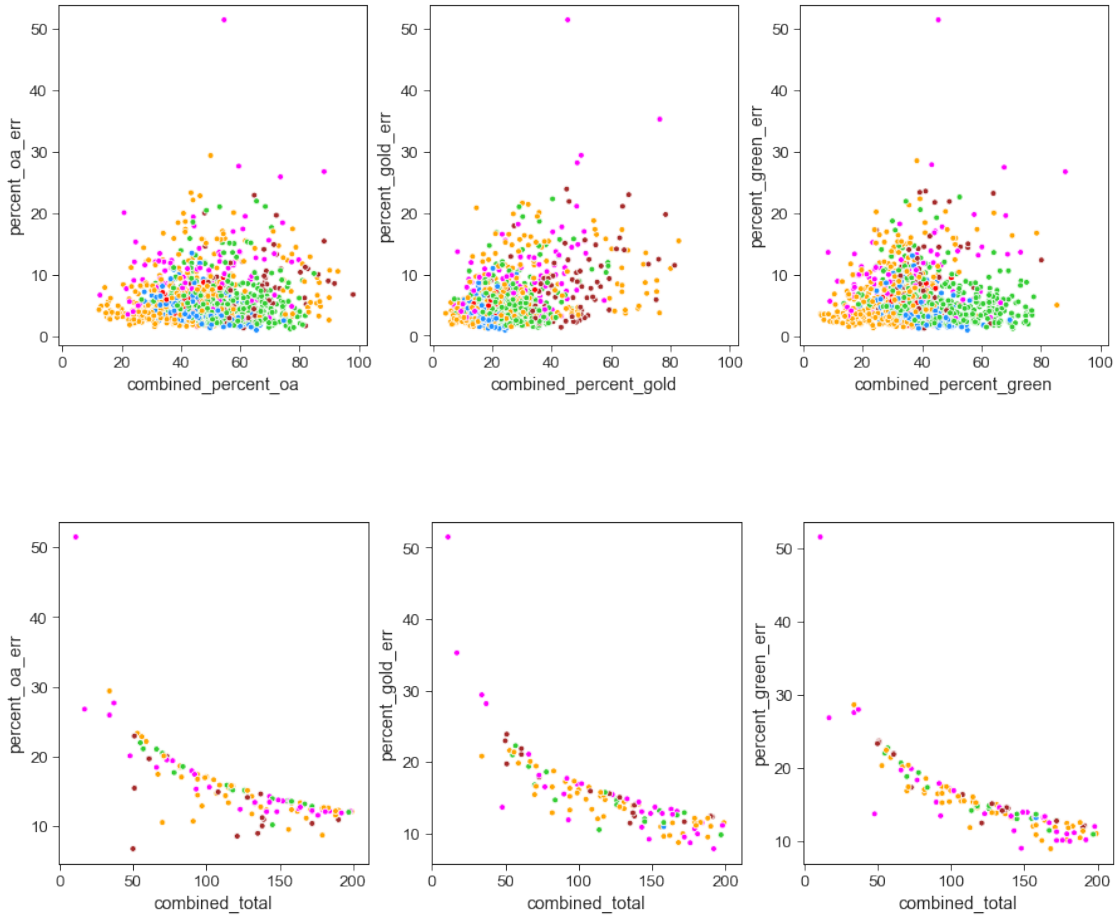


Figure 30: 99.95% Margins of error versus open access percentages and the total number of publications for 2017.



[55]:

	id	country	combined_total	\
1085	c6aecf726ca961a34b512618f23f99d4	Indonesia	94	
3717	ccfe57ae78b78222c04089ac23c25be1	Cyprus	94	
9025	fed9d55804c66969746b952bad0bee55	South Africa	92	
2586	7bedd456436bb878ba7f0c8e26dec960	Indonesia	91	
2702	de2440d52822478457b1624a1b4f4b10	Morocco	90	
20884	c2a2f02b8061094ff9ce2cb6e4faafd5	Romania	84	
22053	c85f0269181519320477bb28bb636b6f	Indonesia	83	
15688	5c5dc64e7d5e3402cbd4f160757a5f39	Jordan	82	
6516	2f8c4b1fd5e24fd132d0b604480f1e75	Philippines	82	
17187	551f380a003d9e6cc7f0ec7f839f8b5d	Lithuania	78	
15802	7e7436071c68567da8ff2e72fb206e71	South Africa	77	
8778	fbcaff06657bef443075e4f099274600	Nigeria	73	
225	9b13795d6e093c1939509ffb68e2156e	Argentina	73	
21990	e86c007680e80afc9d04f3d0395c8ed7	Jordan	71	
21651	3d8ea79a00c683b80ab12d091f9805a2	Lebanon	71	

```
674    ccb509a52635e3cb3aca66ce7a9c5818          Iraq             70
2113   b5b8bc5351921e4cfd35e943c8e5aa74        Latvia             70
15295  7243587f73e82a7fd08333fa7f458f54          Iraq             67
16772  3edaf190bad4326ffacb6c438893d65c         Kenya             66

       published_year  combined_percent_oa  percent_oa_err  \
1085             2017            78.723404        8.273612
3717             2017            41.489362        9.960427
9025             2017            60.869565        9.972857
2586             2017            90.109890        6.133695
2702             2017            44.444444       10.266143
20884            2017            44.047619       10.616636
22053            2017            71.084337        9.753726
15688            2017            41.463415       10.663392
6516             2017            39.024390       10.558331
17187            2017            70.512821       10.119509
15802            2017            44.155844       11.091586
8778             2017            61.643836       11.154678
225              2017            47.945205       11.460343
21990            2017            57.746479       11.490028
21651            2017            40.845070       11.433842
674              2017            92.857143        6.033241
2113             2017            48.571429       11.708458
15295            2017            77.611940        9.981393
16772            2017            74.242424       10.550248

       combined_percent_gold  percent_gold_err  combined_percent_green  \
1085                30.851064          9.337271               58.510638
3717                24.468085          8.690753               31.914894
9025                43.478261         10.129920               46.739130
2586                63.736264          9.877907               70.329670
2702                24.444444          8.878878               23.333333
20884               19.047619          8.397531               30.952381
22053               54.216867         10.718573               40.963855
15688               25.609756          9.447347               28.048780
6516                13.414634          7.376678               25.609756
17187               35.897436         10.645797               44.871795
15802               23.376623          9.453274               32.467532
8778                28.767123         10.384440               57.534247
225                 27.397260         10.231156               24.657534
21990               35.211268         11.110087               47.887324
21651               21.126761          9.495289               23.943662
674                 82.857143          8.829043               78.571429
2113                21.428571          9.612492               30.000000
15295               35.820896         11.481106               64.179104
16772               48.484848         12.057426               68.181818
```

```
      percent_green_err
1085           9.960427
3717           9.423558
9025          10.195455
2586           9.385682
2702           8.738291
20884          9.886392
22053         10.579775
15688          9.723558
6516           9.447347
17187         11.037797
15802         10.459036
8778          11.339066
225            9.887566
21990         11.620073
21651          9.926357
674            9.612492
2113          10.735362
15295         11.481106
16772         11.237147
```

[56]:

```
                                    id       country  combined_total  \
12078  de70d3c3f195c6a82f4dfa3c359dff64     Singapore              94
6160   0bafa5299d607425819728ee2de46faf         Kenya              99
17446  86849c3c648e06a90604e21a68e68fcc        Jordan              99
11089  7d7474c189a5b171f4762f8fa9c3746b         India             100
13838  df1f28e09830fa827503b8e15951b87e  South Africa             102
20765  1f72407fb1213b43e00895101ae5a325          Iraq             105
17143  d63c7168bac203b11383b2cda23d497b     Indonesia             107
17780  66409397d79af215e90e397c18ca9bca      Colombia             108
18772  8d47c1345eb2a75c6c657a2c0dcaf747        Kuwait             109
5023   efac1cc974cac2c09313b2c007dc3a0e        Jordan             113
12958  981a8c9ceb4dbff8493e9982e402945a       Romania             114
3413   d7e5fe1f1ec6c219609e618921a93ebc       Lebanon             115
6239   582d741d9437ff89629340a5a1117239          Iraq             115
20779  2661a078bec2ecfd6e2cca76aa383d35     Indonesia             115
18143  38667f3e1fa24a3fbc69eb481d047002        Cyprus             117
6751   59210822f290ad04e0693c02d8e2c0dd       Romania             117
13769  2f3dacb1b6fc22c24b4cde5933e87a7c          Iraq             117
12206  0a2df6b3b2d45adfa4fe27ca6a620c6c          Iraq             118
13888  b9662aee5b78556e1f204bce9183db9e      Slovakia             118

       published_year  combined_percent_oa  percent_oa_err  \
12078            2017            34.042553        9.579330
6160             2017            57.575758        9.735659
17446            2017            41.414141        9.703071
11089            2017            54.000000        9.768590
```

```
13838            2017           69.607843       8.926184
20765            2017           54.285714       9.528624
17143            2017           72.897196       8.422222
17780            2017           71.296296       8.531918
18772            2017           37.614679       9.094162
5023             2017           27.433628       8.226710
12958            2017           42.982456       9.087690
3413             2017           61.739130       8.883106
6239             2017           54.782609       9.096645
20779            2017           77.391304       7.645238
18143            2017           47.008547       9.043873
6751             2017           62.393162       8.777383
13769            2017           52.136752       9.051826
12206            2017           53.389831       9.000874
13888            2017           63.559322       8.683562

       combined_percent_gold  percent_gold_err  combined_percent_green  \
12078              17.021277          7.597515               24.468085
6160               40.404040          9.666278               44.444444
17446              29.292929          8.965021               27.272727
11089              33.000000          9.216169               41.000000
13838              50.980392          9.701580               44.117647
20765              26.666667          8.458561               36.190476
17143              55.140187          9.423821               31.775701
17780              62.962963          9.107620               42.592593
18772              21.100917          7.660012               33.027523
5023               16.814159          6.895712               15.929204
12958              12.280702          6.025080               26.315789
3413               34.782609          8.705023               51.304348
6239               34.782609          8.705023               35.652174
20779              58.260870          9.012958               50.434783
18143              18.803419          7.080280               35.897436
6751               42.735043          8.963955               30.769231
13769              37.606838          8.777383               33.333333
12206              32.203390          8.430821               34.745763
13888              47.457627          9.009961               32.203390

       percent_green_err
12078           8.690753
6160            9.788383
17446           8.773066
11089           9.639933
13838           9.636060
20765           9.191816
17143           8.822289
17780           9.325995
18772           8.829356
```

```
5023           6.747400
12958          8.083493
3413           9.135437
6239           8.754213
20779          9.138202
18143          8.692257
6751           8.363172
13769          8.541947
12206          8.591527
13888          8.430821
```

[57]:

```
                                    id       country  combined_total  \
6160   0bafa5299d607425819728ee2de46faf         Kenya              99
22097  9932bbdb3ac54518a88edfb037daca20     Indonesia              97
3717   ccfe57ae78b78222c04089ac23c25be1        Cyprus              94
1085   c6aecf726ca961a34b512618f23f99d4     Indonesia              94
9025   fed9d55804c66969746b952bad0bee55  South Africa              92
2586   7bedd456436bb878ba7f0c8e26dec960     Indonesia              91
2702   de2440d52822478457b1624a1b4f4b10       Morocco              90
20884  c2a2f02b8061094ff9ce2cb6e4faafd5       Romania              84
22053  c85f0269181519320477bb28bb636b6f     Indonesia              83
6516   2f8c4b1fd5e24fd132d0b604480f1e75   Philippines              82
15688  5c5dc64e7d5e3402cbd4f160757a5f39        Jordan              82
17187  551f380a003d9e6cc7f0ec7f839f8b5d     Lithuania              78
15802  7e7436071c68567da8ff2e72fb206e71  South Africa              77
225    9b13795d6e093c1939509ffb68e2156e     Argentina              73
8778   fbcaff06657bef443075e4f099274600       Nigeria              73
21651  3d8ea79a00c683b80ab12d091f9805a2       Lebanon              71
21990  e86c007680e80afc9d04f3d0395c8ed7        Jordan              71
674    ccb509a52635e3cb3aca66ce7a9c5818          Iraq              70
2113   b5b8bc5351921e4cfd35e943c8e5aa74        Latvia              70

       published_year  combined_percent_oa  percent_oa_err  \
6160             2017            57.575758       12.795438
22097            2017            83.505155        9.707128
3717             2017            41.489362       13.090846
1085             2017            78.723404       10.873890
9025             2017            60.869565       13.107184
2586             2017            90.109890        8.061427
2702             2017            44.444444       13.492645
20884            2017            44.047619       13.953293
22053            2017            71.084337       12.819183
6516             2017            39.024390       13.876664
15688            2017            41.463415       14.014744
17187            2017            70.512821       13.299926
15802            2017            44.155844       14.577514
225              2017            47.945205       15.062165
```

```
8778             2017           61.643836       14.660434
21651            2017           40.845070       15.027335
21990            2017           57.746479       15.101180
674              2017           92.857143        7.929403
2113             2017           48.571429       15.388260

       combined_percent_gold  percent_gold_err  combined_percent_green  \
6160                40.404040         12.704251               44.444444
22097               54.639175         13.021246               48.453608
3717                24.468085         11.422132               31.914894
1085                30.851064         12.271842               58.510638
9025                43.478261         13.313609               46.739130
2586                63.736264         12.982393               70.329670
2702                24.444444         11.669383               23.333333
20884               19.047619         11.036754               30.952381
22053               54.216867         14.087268               40.963855
6516                13.414634          9.695062               25.609756
15688               25.609756         12.416513               28.048780
17187               35.897436         13.991619               44.871795
15802               23.376623         12.424303               32.467532
225                 27.397260         13.446662               24.657534
8778                28.767123         13.648122               57.534247
21651               21.126761         12.479523               23.943662
21990               35.211268         14.601829               47.887324
674                 82.857143         11.603885               78.571429
2113                21.428571         12.633561               30.000000

       percent_green_err
6160           12.864732
22097          13.071403
3717           12.385248
1085           13.090846
9025           13.399740
2586           12.335468
2702           11.484611
20884          12.993543
22053          13.904847
6516           12.416513
15688          12.779534
17187          14.506819
15802          13.746162
225            12.995086
8778           14.902772
21651          13.046069
21990          15.272096
674            12.633561
2113           14.109333
```

[58]:

	id	country	combined_total	\
2457	7f0587a756359ca7868017a33ce96d55	Algeria	93	
12078	de70d3c3f195c6a82f4dfa3c359dff64	Singapore	94	
17446	86849c3c648e06a90604e21a68e68fcc	Jordan	99	
13838	df1f28e09830fa827503b8e15951b87e	South Africa	102	
20765	1f72407fb1213b43e00895101ae5a325	Iraq	105	
17143	d63c7168bac203b11383b2cda23d497b	Indonesia	107	
17780	66409397d79af215e90e397c18ca9bca	Colombia	108	
18772	8d47c1345eb2a75c6c657a2c0dcaf747	Kuwait	109	
5023	efac1cc974cac2c09313b2c007dc3a0e	Jordan	113	
12958	981a8c9ceb4dbff8493e9982e402945a	Romania	114	
6239	582d741d9437ff89629340a5a1117239	Iraq	115	
20779	2661a078bec2ecfd6e2cca76aa383d35	Indonesia	115	
3413	d7e5fe1f1ec6c219609e618921a93ebc	Lebanon	115	
18143	38667f3e1fa24a3fbc69eb481d047002	Cyprus	117	
6751	59210822f290ad04e0693c02d8e2c0dd	Romania	117	
13769	2f3dacb1b6fc22c24b4cde5933e87a7c	Iraq	117	
13888	b9662aee5b78556e1f204bce9183db9e	Slovakia	118	
7818	52af7c6fcc1f787a4155697d0bd52775	Oman	118	
12206	0a2df6b3b2d45adfa4fe27ca6a620c6c	Iraq	118	
8055	0660c8e502fc0390af8b8d93d5fa70cf	Peru	121	

	published_year	combined_percent_oa	percent_oa_err	\
2457	2017	24.731183	11.524826	
12078	2017	34.042553	12.589976	
17446	2017	41.414141	12.752607	
13838	2017	69.607843	11.731556	
20765	2017	54.285714	12.523334	
17143	2017	72.897196	11.069207	
17780	2017	71.296296	11.213377	
18772	2017	37.614679	11.952328	
5023	2017	27.433628	10.812247	
12958	2017	42.982456	11.943821	
6239	2017	54.782609	11.955591	
20779	2017	77.391304	10.048027	
3413	2017	61.739130	11.674939	
18143	2017	47.008547	11.886233	
6751	2017	62.393162	11.535989	
13769	2017	52.136752	11.896686	
13888	2017	63.559322	11.412682	
7818	2017	47.457627	11.841663	
12206	2017	53.389831	11.829720	
8055	2017	91.735537	6.448063	

	combined_percent_gold	percent_gold_err	combined_percent_green	\
2457	12.903226	8.954773	17.204301	
12078	17.021277	9.985306	24.468085	

```
17446              29.292929         11.782599              27.272727
13838              50.980392         12.750649              44.117647
20765              26.666667         11.116966              36.190476
17143              55.140187         12.385593              31.775701
17780              62.962963         11.970015              42.592593
18772              21.100917         10.067445              33.027523
5023               16.814159          9.062936              15.929204
12958              12.280702          7.918676              26.315789
6239               34.782609         11.440888              35.652174
20779              58.260870         11.845601              50.434783
3413               34.782609         11.440888              51.304348
18143              18.803419          9.305511              35.897436
6751               42.735043         11.781198              30.769231
13769              37.606838         11.535989              33.333333
13888              47.457627         11.841663              32.203390
7818               38.983051         11.565596              30.508475
12206              32.203390         11.080507              34.745763
8055               51.239669         11.705491              80.165289

       percent_green_err
2457           10.081538
12078          11.422132
17446          11.530316
13838          12.664536
20765          12.080672
17143          11.595009
17780          12.257023
18772          11.604296
5023            8.868012
12958          10.624019
6239           11.505537
20779          12.010208
3413           12.006574
18143          11.424109
6751           10.991598
13769          11.226559
13888          11.080507
7818           10.918954
12206          11.291721
8055            9.338117
```

[59]:

```
                                    id       country  combined_total  \
11089  7d7474c189a5b171f4762f8fa9c3746b         India             100
6160   0bafa5299d607425819728ee2de46faf         Kenya              99
22097  9932bbdb3ac54518a88edfb037daca20     Indonesia              97
3717   ccfe57ae78b78222c04089ac23c25be1        Cyprus              94
1085   c6aecf726ca961a34b512618f23f99d4     Indonesia              94
```

```
9025   fed9d55804c66969746b952bad0bee55  South Africa             92
2586   7bedd456436bb878ba7f0c8e26dec960     Indonesia             91
2702   de2440d52822478457b1624a1b4f4b10       Morocco             90
20884  c2a2f02b8061094ff9ce2cb6e4faafd5       Romania             84
22053  c85f0269181519320477bb28bb636b6f     Indonesia             83

       published_year  combined_percent_oa  percent_oa_err  \
11089            2017            54.000000       17.095032
6160             2017            57.575758       17.037403
22097            2017            83.505155       12.925252
3717             2017            41.489362       17.430747
1085             2017            78.723404       14.478821
9025             2017            60.869565       17.452500
2586             2017            90.109890       10.733966
2702             2017            44.444444       17.965750
20884            2017            44.047619       18.579113
22053            2017            71.084337       17.069021

       combined_percent_gold  percent_gold_err  combined_percent_green  \
11089              33.000000         16.128296               41.000000
6160               40.404040         16.915986               44.444444
22097              54.639175         17.338072               48.453608
3717               24.468085         15.208817               31.914894
1085               30.851064         16.340224               58.510638
9025               43.478261         17.727359               46.739130
2586               63.736264         17.286338               70.329670
2702               24.444444         15.538037               23.333333
20884              19.047619         14.695678               30.952381
22053              54.216867         18.757504               40.963855

       percent_green_err
11089          16.869882
6160           17.129671
22097          17.404857
3717           16.491226
1085           17.430747
9025           17.842046
2586           16.424944
2702           15.292009
20884          17.301185
22053          18.514606
```

[60]:

```
                                    id       country  combined_total  \
2457   7f0587a756359ca7868017a33ce96d55       Algeria              93
12078  de70d3c3f195c6a82f4dfa3c359dff64     Singapore              94
17446  86849c3c648e06a90604e21a68e68fcc        Jordan              99
13838  df1f28e09830fa827503b8e15951b87e  South Africa             102
```

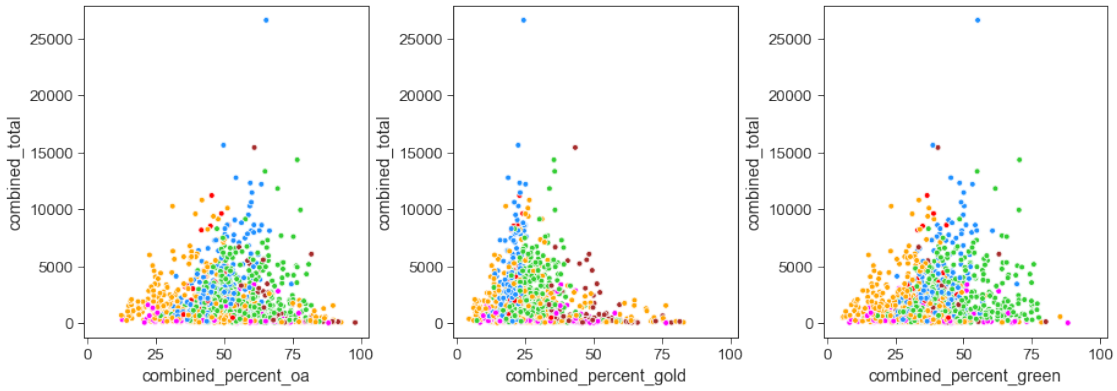
20765	1f72407fb1213b43e00895101ae5a325	Iraq	105
17143	d63c7168bac203b11383b2cda23d497b	Indonesia	107
17780	66409397d79af215e90e397c18ca9bca	Colombia	108
18772	8d47c1345eb2a75c6c657a2c0dcaf747	Kuwait	109
5023	efac1cc974cac2c09313b2c007dc3a0e	Jordan	113
12958	981a8c9ceb4dbff8493e9982e402945a	Romania	114
6239	582d741d9437ff89629340a5a1117239	Iraq	115
20779	2661a078bec2ecfd6e2cca76aa383d35	Indonesia	115
3413	d7e5fe1f1ec6c219609e618921a93ebc	Lebanon	115
18143	38667f3e1fa24a3fbc69eb481d047002	Cyprus	117
6751	59210822f290ad04e0693c02d8e2c0dd	Romania	117
13769	2f3dacb1b6fc22c24b4cde5933e87a7c	Iraq	117
13888	b9662aee5b78556e1f204bce9183db9e	Slovakia	118
7818	52af7c6fcc1f787a4155697d0bd52775	Oman	118
12206	0a2df6b3b2d45adfa4fe27ca6a620c6c	Iraq	118
8055	0660c8e502fc0390af8b8d93d5fa70cf	Peru	121

	published_year	combined_percent_oa	percent_oa_err	\
2457	2017	24.731183	15.345556	
12078	2017	34.042553	16.763827	
17446	2017	41.414141	16.980374	
13838	2017	69.607843	15.620822	
20765	2017	54.285714	16.675091	
17143	2017	72.897196	14.738889	
17780	2017	71.296296	14.930856	
18772	2017	37.614679	15.914784	
5023	2017	27.433628	14.396742	
12958	2017	42.982456	15.903458	
6239	2017	54.782609	15.919129	
20779	2017	77.391304	13.379166	
3413	2017	61.739130	15.545435	
18143	2017	47.008547	15.826778	
6751	2017	62.393162	15.360420	
13769	2017	52.136752	15.840696	
13888	2017	63.559322	15.196234	
7818	2017	47.457627	15.767432	
12206	2017	53.389831	15.751529	
8055	2017	91.735537	8.585736	

	combined_percent_gold	percent_gold_err	combined_percent_green	\
2457	12.903226	11.923475	17.204301	
12078	17.021277	13.295652	24.468085	
17446	29.292929	15.688786	27.272727	
13838	50.980392	16.977766	44.117647	
20765	26.666667	14.802482	36.190476	
17143	55.140187	16.491687	31.775701	
17780	62.962963	15.938335	42.592593	

18772	21.100917	13.405022	33.027523
5023	16.814159	12.067497	15.929204
12958	12.280702	10.543890	26.315789
6239	34.782609	15.233791	35.652174
20779	58.260870	15.772676	50.434783
3413	34.782609	15.233791	51.304348
18143	18.803419	12.390490	35.897436
6751	42.735043	15.686922	30.769231
13769	37.606838	15.360420	33.333333
13888	47.457627	15.767432	32.203390
7818	38.983051	15.399842	30.508475
12206	32.203390	14.753936	34.745763
8055	51.239669	15.586116	80.165289

	percent_green_err
2457	13.423787
12078	15.208817
17446	15.352866
13838	16.863105
20765	16.085678
17143	15.439006
17780	16.320492
18772	15.451372
5023	11.807950
12958	14.146113
6239	15.319872
20779	15.991853
3413	15.987015
18143	15.211450
6751	14.635551
13769	14.948408
13888	14.753936
7818	14.538824
12206	15.035172
8055	12.433905



[62]:

	id	country	country_code	\
9140	17a0a6aa32cb509bdec28ee32b92087d	Uganda	UGA	
17201	75f0b05294ac39ed2e42f59866109298	Kenya	KEN	
19306	fc32c0bb1e2a586a23d3a6895162ae04	Philippines	PHL	
8144	7f3f85f9a081644c3b3b809688956b08	Kenya	KEN	
17425	56a76c9358e9849e258db85162159e00	South Africa	ZAF	
…	…	…	…	
13045	860d3987e4ee679f50165a2e108403c1	Russian Federation	RUS	
701	c207a92177830deff6ab32ea020c1aa1	Indonesia	IDN	
5218	fb23ff4dd57689d52cc4298180631a81	Morocco	MAR	
5337	c3206e1a4f8743bcbc08c9c44c603fc8	Iraq	IRQ	
11874	1365e23c734d32b553da973c58a32a58	Jordan	JOR	

	region	subregion	published_year	combined_total	\
9140	Africa	Sub-Saharan Africa	2017	11	
17201	Africa	Sub-Saharan Africa	2017	17	
19306	Asia	South-eastern Asia	2017	34	
8144	Africa	Sub-Saharan Africa	2017	34	
17425	Africa	Sub-Saharan Africa	2017	37	
…	…	…	…	…	
13045	Europe	Eastern Europe	2017	172	
701	Asia	South-eastern Asia	2017	179	
5218	Africa	Northern Africa	2017	179	
5337	Asia	Western Asia	2017	179	
11874	Asia	Western Asia	2017	180	

	combined_oa	combined_green	combined_gold	…	just_mag_percent_oa	\
9140	6	5	5	…	62.500000	
17201	15	15	13	…	NaN	
19306	17	13	5	…	51.515152	
8144	25	23	17	…	NaN	
17425	22	16	18	…	51.612903	
…	…	…	…	…	…	
13045	101	68	73	…	29.310345	
701	155	35	127	…	89.024390	
5218	66	43	40	…	NaN	
5337	105	67	61	…	58.870968	
11874	79	55	46	…	37.681159	

	just_mag_percent_green	just_mag_percent_gold	just_mag_percent_hybrid	\
9140	62.500000	62.500000	12.500000	
17201	NaN	NaN	NaN	
19306	39.393939	15.151515	6.060606	
8144	NaN	NaN	NaN	
17425	35.483871	41.935484	12.903226	

```
…                      …                      …                      …
13045               17.241379              20.689655               1.724138
701                 20.121951              74.390244              18.292683
5218                      NaN                    NaN                    NaN
5337                37.903226              32.258065              16.129032
11874               23.188406              25.362319               7.971014

       just_mag_percent_bronze  just_mag_percent_green_only  \
9140                  0.000000                     0.000000
17201                      NaN                          NaN
19306                21.212121                    15.151515
8144                       NaN                          NaN
17425                 9.677419                     0.000000
…                          …                            …
13045                 0.000000                     8.620690
701                  12.804878                     1.829268
5218                       NaN                          NaN
5337                 12.096774                    14.516129
11874                 7.246377                     5.072464

       just_mag_percent_in_home_repo  percent_oa_err  percent_gold_err  \
9140                             0.0       51.495078         51.495078
17201                            0.0       26.802874         35.287623
19306                            0.0       29.412007         20.833380
8144                             0.0       25.951771         29.412007
17425                            0.0       27.685253         28.184127
…                                …                …                 …
13045                            0.0       12.876307         12.926486
701                              0.0        8.735465         11.639065
5218                             0.0       12.368749         10.679526
5337                             0.0       12.624817         12.151237
11874                            0.0       12.687019         11.151073

       percent_green_err
9140           51.495078
17201          26.802874
19306          28.586242
8144           27.519213
17425          27.935804
…                    …
13045          12.787115
701            10.167867
5218           10.952626
5337           12.406834
11874          11.776649

[100 rows x 69 columns]
```

[63]:

	id	country	combined_total	\
8055	0660c8e502fc0390af8b8d93d5fa70cf	Peru	121	
12451	aa85927a24ec106456a11e9be6c427b9	Colombia	135	
2567	1f78c981ce195c43f3a54e68db290f30	Indonesia	156	
701	c207a92177830deff6ab32ea020c1aa1	Indonesia	179	
13794	7917c4608f10135692aee554ab9df5fb	Croatia	220	
…	…	…	…	
456	fb6c522904e94945a8840f20c6a893dd	Philippines	404	
6404	80f21c8e839d66f832c4b43dbf69cd4c	Austria	406	
20622	f8541681a140edcb0e101f59852b39bb	Malaysia	407	
11889	41c97a3e44f22c74c7c34a959cfc0fca	Slovakia	407	
8224	4bedb3086a6d92161a9d5b1ab427bc20	Saudi Arabia	408	

	published_year	combined_percent_oa	percent_oa_err
8055	2017	91.735537	8.585736
12451	2017	89.629630	9.000159
2567	2017	85.897436	9.558087
701	2017	86.592179	8.735465
13794	2017	80.909091	9.088553
…	…	…	…
456	2017	31.683168	7.939289
6404	2017	32.512315	7.973836
20622	2017	36.363636	8.178681
11889	2017	44.717445	8.453362
8224	2017	58.333333	8.371761

[100 rows x 6 columns]

[64]:

	id	country	\
4711	5890cd84c0d837426db43eab59fc073b	South Africa	
11336	ed088c3da771a00d7bce9277a89b3233	United Kingdom	
1921	3aa2e52f2e97b9a908ae1371aabb0584	Saudi Arabia	
14898	1b7b2deaba20df343f6f791b4a611147	South Africa	
7670	90b6459198c9ce1244e8958863f7e0f0	United States of America	
…	…	…	
11133	85ae3d316ecca8aecb0611c349122270	Lithuania	
1952	6ae049db63d0cba0afff62efad859f33	Lithuania	
1208	8dcf5e1f6d7e6eae74daa7120abd5c55	Uganda	
7818	52af7c6fcc1f787a4155697d0bd52775	Oman	
13888	b9662aee5b78556e1f204bce9183db9e	Slovakia	

	combined_total	published_year	combined_percent_oa	percent_oa_err
4711	290	2017	49.655172	10.070587
11336	286	2017	54.895105	10.092291
1921	284	2017	57.042254	10.075208
14898	282	2017	54.609929	10.169177
7670	280	2017	39.285714	10.011010

```
…              …              …                 …              …
11133          126            2017              56.349206      15.154752
1952           126            2017              57.936508      15.084734
1208           123            2017              79.674797      12.445705
7818           118            2017              47.457627      15.767432
13888          118            2017              63.559322      15.196234

[100 rows x 6 columns]
```

[65]:

```
                                    id      country  combined_total  \
10541  c9205c13f40f38dea019321bd31e1934      Algeria             333
17974  517407018f965e8b374beb4daddd8519    Singapore             368
2601   9f4b9655846320936752c72df50211e3        India             418
9758   a64ea852899d89c7363e20cab18cf5db      Denmark             447
18320  36cb9efa50f21c7a87302a7944be8768        India             448
…                                   …            …              …
12128  7fff6e3e880547cfdb3de19b872e755f      Germany            1009
4275   1668ec5ca2800f794a11e0b1a8c5daed        India            1010
15733  8455f542015081dd933fc5ee5e863095  New Zealand            1010
13560  b2ecd23a19a5193a49b1bc07c56d19a7        China            1010
12773  45a04b85ec0d8cdf9cc7ed12681b5c05       Greece            1017

       published_year  combined_percent_gold  percent_gold_err
10541            2017               7.207207          4.860855
17974            2017               4.347826          3.646311
2601             2017               6.220096          4.051906
9758             2017               8.277405          4.470186
18320            2017               8.258929          4.460657
…                …                        …                 …
12128            2017              26.759167          4.780370
4275             2017              21.683168          4.447566
15733            2017              18.316832          4.174696
13560            2017              17.920792          4.139316
12773            2017              25.663717          4.697793

[100 rows x 6 columns]
```

[66]:

```
                                    id                  country  \
14351  7b3196a4420c2549c81e82948a551154                   Taiwan
8689   d1556fc44d6a31d6d7301e48c760219f                  Hungary
10288  7ec944f82f99a3d8a88d372fa4f57752                    Japan
13131  90a871b73d1c5f5f3c48ad259cf1677c             South Africa
2219   ddbba79fdb43895d86a11eb3cbf19950       Russian Federation
…                                   …                        …
1999   e0f78234c1f108a03bf329b7bb8eac01                    Japan
11229  c4d614fec3cc12993d32b78575787527           United Kingdom
18592  457d3f1a129f8983cfa53ef5bd272d69                   Brazil
```

```
2203   31f7038d6ed7cee70d9608366f108a42                    Italy
15928  19835444d4bded5809a5fc79d4722eea  United States of America

       combined_total  published_year  combined_percent_gold  percent_gold_err
14351            1014            2017              32.445759          5.042901
8689             1011            2017              33.036597          5.073819
10288            1010            2017              31.485149          5.012780
13131            1001            2017              31.368631          5.030211
2219              975            2017              38.256410          5.338753
…                 …               …                      …                 …
1999              588            2017              30.102041          6.488371
11229             587            2017              33.219761          6.668026
18592             586            2017              38.907850          6.908068
2203              586            2017              15.017065          5.061785
15928             585            2017              23.418803          6.005644

[100 rows x 6 columns]
```

[67]:

```
                                  id  country  combined_total  \
2262   02ae42f3a09c296557a2064bda0a13ad     Iran             480
4997   2820f15cfb23fba4dce34bd1414c0101    India             557
9104   292f00c7a6541e39c2e6b5003ec55198     Iran             583
21553  15f33f12c929f86f765800b4ed17ab30  Algeria             649
20439  55111cbefb3e539dd79c13a1add179b9    India             686
…                                   …        …               …
6552   a49e79754beacd0d386bf2bb9ef398f6     Iran            1332
9799   dad0f67ec4ed1d5f3e1a136b221d43bf  Finland            1332
7962   e68338b4276f7bfa08ee2e5495482691    Japan            1345
11979  0b6a437c6f64e4a19b1a1134ad16bb3c  Czechia            1356
1424   63edb9eb09980c0957b79938dd90ae08   Turkey            1357

       published_year  combined_percent_green  percent_green_err
2262             2017                5.833333          3.669277
4997             2017                6.104129          3.479378
9104             2017                7.032590          3.632310
21553            2017               14.791988          4.779971
20439            2017                8.454810          3.643357
…                 …                       …                 …
6552             2017               17.717718          3.588385
9799             2017               63.888889          4.514141
7962             2017               30.037175          4.287422
11979            2017               31.637168          4.331847
1424             2017               24.023581          3.977979

[100 rows x 6 columns]
```

[68]:

	id	country	combined_total	\
14906	7a44d402c8b0cd832f62bdaaff487639	United Kingdom	1170	
5715	756dbbeb1ad66252febb228a6ad22c7c	Argentina	1155	
11651	b79a7ec6eb08f77037ec649ebe2751e1	Spain	1142	
16027	ec9e8837955f09335824cb57905badea	Japan	1135	
9861	7ae337d4f871620f47b2d7deff0d65d8	Hungary	1130	
…	…	…	…	
13783	c5c56fe5509ce646b308174ab4e4e587	Indonesia	764	
11080	078f4d45c3c9cdbea0f3054bd4290e42	Colombia	764	
16594	bc9acc9459162a49a73dbd3d6de8d63e	South Korea	757	
19487	8de10a217baa695607525f7c5b6e9137	Australia	755	
20846	9cc802b47b2dec063a68b13196db713c	Australia	751	

	published_year	combined_percent_green	percent_green_err
14906	2017	49.401709	5.013489
5715	2017	43.809524	5.007475
11651	2017	41.593695	5.002703
16027	2017	43.964758	5.053347
9861	2017	54.070796	5.084880
…	…	…	…
13783	2017	32.068063	5.791897
11080	2017	34.162304	5.885161
16594	2017	28.533686	5.629578
19487	2017	33.509934	5.892313
20846	2017	29.693742	5.718783

[100 rows x 6 columns]

[69]:

	id	country	\
14906	7a44d402c8b0cd832f62bdaaff487639	United Kingdom	
11760	8692a1b46cc982a13d56bbda20f5ba0d	Japan	
513	5743617c036a3a3520f9fac51de55212	Czechia	
5715	756dbbeb1ad66252febb228a6ad22c7c	Argentina	
11651	b79a7ec6eb08f77037ec649ebe2751e1	Spain	
…	…	…	
22391	9c71d9ccc841663780c9fdec33b8b9e5	Spain	
18113	33db425226116d4bf2f135e0042f0113	Pakistan	
12014	50bad8379813e2c57dbd1f60dd02875d	Ireland	
4292	b4dab81a017f4636a80b13ea658333a4	United States of America	
14494	cb5bc7025c1cdf2e0392b22d5706dcf9	Costa Rica	

	combined_total	published_year	combined_percent_oa	percent_oa_err	\
14906	1170	2017	55.897436	4.978850	
11760	1168	2017	47.003425	5.009119	
513	1163	2017	47.721410	5.023690	
5715	1155	2017	56.103896	5.008557	
11651	1142	2017	52.889667	5.066459	

```
…                …                …                    …                …
22391              860              2017              57.558140         5.780900
18113              855              2017              33.450292         5.534573
12014              851              2017              50.411281         5.878745
4292               846              2017              61.583924         5.735868
14494              846              2017              70.921986         5.355275

       combined_percent_gold  percent_gold_err  combined_percent_green  \
14906              24.700855          4.324662               49.401709
11760              23.030822          4.225575               36.130137
513                28.202923          4.525897               37.661221
5715               31.515152          4.688782               43.809524
11651              22.329247          4.226948               41.593695
…                        …                …                       …
22391              25.813953          5.118389               48.837209
18113              18.479532          4.552919               25.614035
12014              16.921269          4.408499               46.415981
4292               25.413712          5.134196               56.028369
14494              54.609929          5.871177               42.671395

       percent_green_err
14906           5.013489
11760           4.821204
513             4.873384
5715            5.007475
11651           5.002703
…                  …
22391           5.846520
18113           5.120302
12014           5.863822
4292            5.853279
14494           5.832611

[100 rows x 10 columns]
```

5 Summary of main findings and implications

This article provided evidence for the sensitivities of comprehensive data workflow proposed in Huang et al. (2020b) against the choices of bibliographic data sources, versions of data sources, use of open access definitions, and confidence levels in statistical inference. Significant differences in the estimated institutional open access levels can arise through choices made on these factors. These differences can also result in biases toward a university's geographical location, preferred choice of open access route, and the time of publication.

This sensitivity analysis is essential for building a robust and fair evaluation framework for open access, and for understanding differences across data sources (in terms of coverage) and university groupings (in terms of regional foci). It also implies that any process for evaluating open access

should clearly describe what data sources are used, which version of data is used, and how they are used in any standardisation and filtering procedures.

6 References

1. Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. PLoS ONE, 5(6), e11273. https://doi.org/10.1371/journal.pone.0011273
2. Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020a). Comparison of bibliographic data sources: Implications for the robustness of university rankings. Quantitative Science Studies, 1(2), 445-478. https://doi.org/10.1162/qss_a_00031
3. Huang, C.-K. (Karl), Neylon, C., Hosking, R., Brookes-Kenworthy, C., Montgomery, L., Wilson, K., & Ozaygen, A. (2020b). Evaluating institutional open access performance: Methodology, challenges and assessment. Under Review, 1–18. https://doi.org/10.5281/zenodo.3694943
4. Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009. PLoS ONE, 6(6), e20961. https://doi.org/10.1371/journal.pone.0020961
5. Matsubayashi, M., Kurata, K., Sakai, Y., Morioka, T., Kato, S., Mine, S., & Ueda, S. (2009). Status of open access in the biomedical field in 2005. Journal of the Medical Library Association : JMLA, 97(1), 4–11. https://doi.org/10.3163/1536-5050.97.1.002