Measuring the Use of Data in Academia

Lucas Kitzmuller, Daniel Mahler, Umar Serajuddin, Brian Stacy, and Xiaoyu Wang

2022-10-07

A concentration of high-quality academic research can shed light on well-being within a country, shape public policy, and monitor progress on goals. A key ingredient to having an abundance of high-quality research in a country is timely, high quality, and accessible data coming from the national statistical system. However, little has been done to document how well statistical systems are supporting the use of data by academia, and currently no global database exists documenting the use of data by academia for statistical systems. This paper lays out an approach for producing such a metric of data use by academia. Natural Language Processing (a form of text mining) is applied to open-access research papers, from which a measure of data use by academia is produced for each country. Compared to a set of aroun 3,500 human coded academic articles, the natural language processing approach produces country rankings that have a correlation of over 0.96 with the human coded approach. After classifying more than 1 million academic articles, measures of the number of articles using data have a strong correlation between GDP, Population, and the Statistical Performance Indicator (SPI) overall score. Around 50% of all papers using data from 2000-2020 were produced by high income countries, while low income countries only produce around 5% of articles using data.

# Introduction

International donors have recognized the important role that national statistical systems play in producing high quality data related to development, such as in producing censuses, surveys, indicators, and in providing users data services. According to PARIS21, the international community funded around $2 billion in statistical development activities between 2016 and 2018 (Yu Tian (2020)). The World Bank alone has funded $610 million for those activities in that period. These investments are typically justified as a means to improve evidence-based policy making in low- and middle-income countries.

Academia plays a key role in producing evidence-based policy recommendations and is a strong beneficiary of the existence of productive national statistical systems. High quality population censuses, for instance, are key for producing a representative sample in a survey that is part of a research study. Microdata produced by the national statistical system in the form of household income and expenditure surveys, agriculture surveys, or demographic and health surveys are routinely used as part of academic studies. A dysfunctional system that is not producing the key data sources or indicators, does not follow basic international classifications to make data comparable, or does not have the data services in place to connect users of the data to the products that are produced, may have less engagement by academia and consequently have a deficit in research on problems facing the country.

The World Bank recognized the importance of academia as a user of the national statistical system in the recently released Statistical Performance Indicators (SPI). The SPI framework assesses the maturity and performance of national statistical systems in five key areas referred to as pillars (Dang et al. (2021)). The first pillar is on data use. It considers the extent that a national statistical system produces products that are meeting the needs of users, and one of the dimensions of this pillar is on data use by academia. However, as acknowledged in the report, measures of data use by academia are in an embryonic stage. Without measures highlighting deficiencies in the usage of data, it is difficult to spotlight problem areas, and countries may lack the information needed to prioritize investments in statistics.

Some previous academic research has highlighted gaps between countries in academic research output. For instance, Robinson, Hartley, and Schneider (2006), Das et al. (2013), and Porteous (2020) examine which countries are studied most by economists using the EconLit database. These studies note that the distribution of economic research is highly uneven across countries. Cameron, Mishra, and Brown (2016) and Sabet and Brown (2018) extend this to note that impact evaluations are highly uneven across countries as well. However, these studies are narrowly focused on publication in economics journals and the authors are not updating their rankings across countries on an annual basis. Another issue is the EconLit database contains data for only 74 countries.

Advances in Natural Language Processing, based on machine learning, and new data sources have opened up new approaches to measuring academic data use at low costs. Organizations such as Semantic Scholar Open Research Corpus (s2orc) have digitized millions of research papers produced around the world, including in fields other than economics, and opened up APIs for accessing the raw text of these documents (Lo et al. (2020)). The s2orc dataset also includes data from 216 countries, which is much greater than the 74 countries covered by EconLit.[[1]](#footnote-20) Natural Language Processing algorithms have been developed to efficiently parse millions of documents to classify documents. For instance, Hansen, McMahon, and Prat (2018), published in the Quarterly Journal of Economics, processed 149 transcripts from the Federal Reserve to study how Federal Reserve board members react to transparency in terms of topic coverage.

Using these tools, we propose to develop a methodology for measuring data use by academia. These measures will then be used to study the relationship between products produced by the national statistical system and research output,

The project seeks to answer the following questions:

1. What is a feasible way to create a metric of data use by academia?
2. What is the relationship between activities of the national statistical system and academic output?

# Emprical Strategy

The empirical approach taken in this project is based on text mining using Natural Language Processing (NLP). The goal of NLP is to take raw text with no structure and extract information from that raw text in a structured way. In our case, NLP will be used to extract the country of study and whether the paper utilizes data. Each of these will be discussed in turn.

Before proceeding to a discussion of the NLP techniques a brief discussion will be given of the primary dataset. The primary dataset for this analysis will be the s2orc dataset containing the raw text of over 80 million open access articles. These include working papers as well as published articles. The s2orc team collects open access research outputs based on institutional repositories indexed by Semantic Scholar. This data source will be described in more detail below, but the key is that this data source provides the title and abstract, as well as other metadata information about the paper such as the subject, journal, and citation counts, through an API that facilitates automated access to these articles through a computer script.

First, the country or countries of study in each academic article will be determined by using two approaches based on the information in the title, abstract, or topic fields. The first does regular expression searches based on the presence of ISO3166 country names. For this approach, a defined set of country names are compiled and the presence of these country names is checked within the title, abstract, or topic fields. Advantages of this approach are that it is transparent, widely used in social science research, and extends to other languages easily. A downside of this approach is that there are potential exclusion errors, such as if the country’s spelling is non-standard.

A second approach is based on Named Entity Recognition (NER), which uses machine learning to identify objects from text, using the Python library spaCy. SpaCy has been used by a number of academic studies to classify text, such Kleinberg et al. (2018) & Shelar et al. (2020). The Named Entity Recognition algorithm splits text into a set of named entities. To give an example, the text “Apple is looking at buying U.K. startup for $1 billion”, can be classified as “Apple [ORG] is looking at buying U.K. [Country] startup for $1 billion [Monetary value]”. For the purposes of this study, NER will be used to identify countries of study in the academic articles. The spaCy python package supports multiple languages and has been trained on multiple spellings of countries and so overcomes some of the limitations of the regular expression approach.

If a country shows up either through the regular expression search or the search using NER, the country will be linked to that article. Using this approach, it may also be possible to acquire additional information such as the topic of study (i.e. health, education, trade), which could give further breakdowns of academic data use.

Second, whether the paper makes use of data will be classified. Again two approaches will be tried. First, a regular expression search approach to classify papers will be applied again. In this case, the algorithm will check for the presence of key words, such as “data”, “survey”, “census” in the text of articles. Second, supervised machine learning, where the team creates a precise definition of data use, manually labels 3500 publications randomly selected, and then trains a model on this data will also be explored.

While text mining based on Natural Language Processing has been used successfully in other contexts, such as the Hansen, McMahon, and Prat (2018) analysis of Federal Reserve transcripts, it is not guaranteed that the approach will succeed. In order to determine whether the approaches tested in this paper are successful, the team will manually classify 3500 randomly selected publications based on the country of study and whether the article makes use of data in the analysis. A key research output of this project will be the extent that the classifications of country of study and use of data match those classified manually by the team.

Assuming the approach to classifying the country of study and use of data are successful, three measures of data use by academia will be constructed as follows. These measures will be produced for countries on an annual basis, and a time series will be created based on the year of the publication of the article dating back to year 2002, which is the start of the s2orc dataset time series. The first use of data measure is a per capita measure, a second measure will be a regression adjusted measure that adjust for other factors in the country such as the population size and GDP, and a third measure we will consider is the share of papers on a country that use data.

# Data

The primary data source for academic articles comes from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al. (2020)). The corpus contains more than 80 million English language academic papers across multiple disciplines which contain . The papers included in the Semantic Scholar corpus are gathered directly from publishers, from open archives such as arXiv or PubMed, and crawled from the internet.

The S2orc corpus contains more than 80 million articles. Some restrictions were placed on the articles to make them usable and relevant to measuring whether the statistical system is releasing usable data for academia. First, only articles with an abstract and parsed PDF or latex file are included in the analysis. The full text of the abstract is necessary to classify the country of study and whether the article uses data. The parsed PDF and latex file are important for extracting important information like the data of publication and field of study. This restriction eliminated a large number of articles in the original corpus Around 30 million articles remain after keeping only articles with a parsable PDF, and around 26% of that 30 million are eliminated when keep articles with an abstract. Second, only articles from the year 2000 to 2020 were considered. This restriction eliminated an additional 9% of articles. Finally, articles from the following fields of study were excluded, as they were unlikely to involve data produced by the national statistical system: Biology, Chemistry, Engineering, Physics, Materials Science, Environmental Science, Geology, History, Philosophy, Math, Computer Science, and Art. Fields that *are* included are: Economics, Political Science, Business, Sociology, Medicine, and Psychology. This third restriction eliminated around 34% of articles. From an initial corpus of 136 million articles, this resulted in a final corpus of around 10 million articles.

A set of 3,500 articles were then randomly selected from this set of 10 million for human classification by raters using the Mechanical Turk service. The raters were chosen from a pool of [Mechanical Turk master workers](https://www.mturk.com/worker/help). A master worker is a rater who has been designated a Mechanical Turk Masters qualification by Amazon. The workers are designated as a master worker on the basis of prior submissions, with a track record of having submitted high quality results, based on approver rates, tenure, and variety of work performed. 137 unique raters were used to classify the articles. The median amount of time that a rater spent on an article, measured as the time between when the article was accepted by the rater and when the rating was submitted was 25.4 minutes. If human raters were used, rather than machine learning tools, then the corpus of 1,037,748 articles examined in this study would take 439,602 hours of human work time to review at a cost of $3,113,244, which assumes a cost of $3 per article as paid to Mturk workers during the team’s article classification exercise described.

Finally, due to the intensive computer resources required, a set of 1,037,748 articles were randomly selected from the 10 million articles in our restricted corpus as a convenience sample. Summary statistics of the final sample of 1 million articles are available in the Table below.

Table. Summary Statistics of Article Corpus. 2000-2020

| Field | Published in Journal (1=yes) | Articles | Share of Articles |
| --- | --- | --- | --- |
| Business | 0.44 | 28,571 | 2.8 |
| Economics | 0.21 | 62,241 | 6.0 |
| Medicine | 0.04 | 840,920 | 81.0 |
| Political Science | 0.58 | 26,185 | 2.5 |
| Psychology | 0.25 | 44,191 | 4.3 |
| Sociology | 0.10 | 35,640 | 3.4 |

# Machine Learning Approach

Add in Xiaoyu’s description.

After training and validating the model, the country (or countries) of study and whether the article article uses data are classified for a corpus of 1,037,748 academic articles. Using this classified dataset of articles, a new dataset was constructed with the number of articles using data per country. Countries are organized to match the countries in the World Bank’s World Development Indicators. In total, article counts are available for 218 countries.

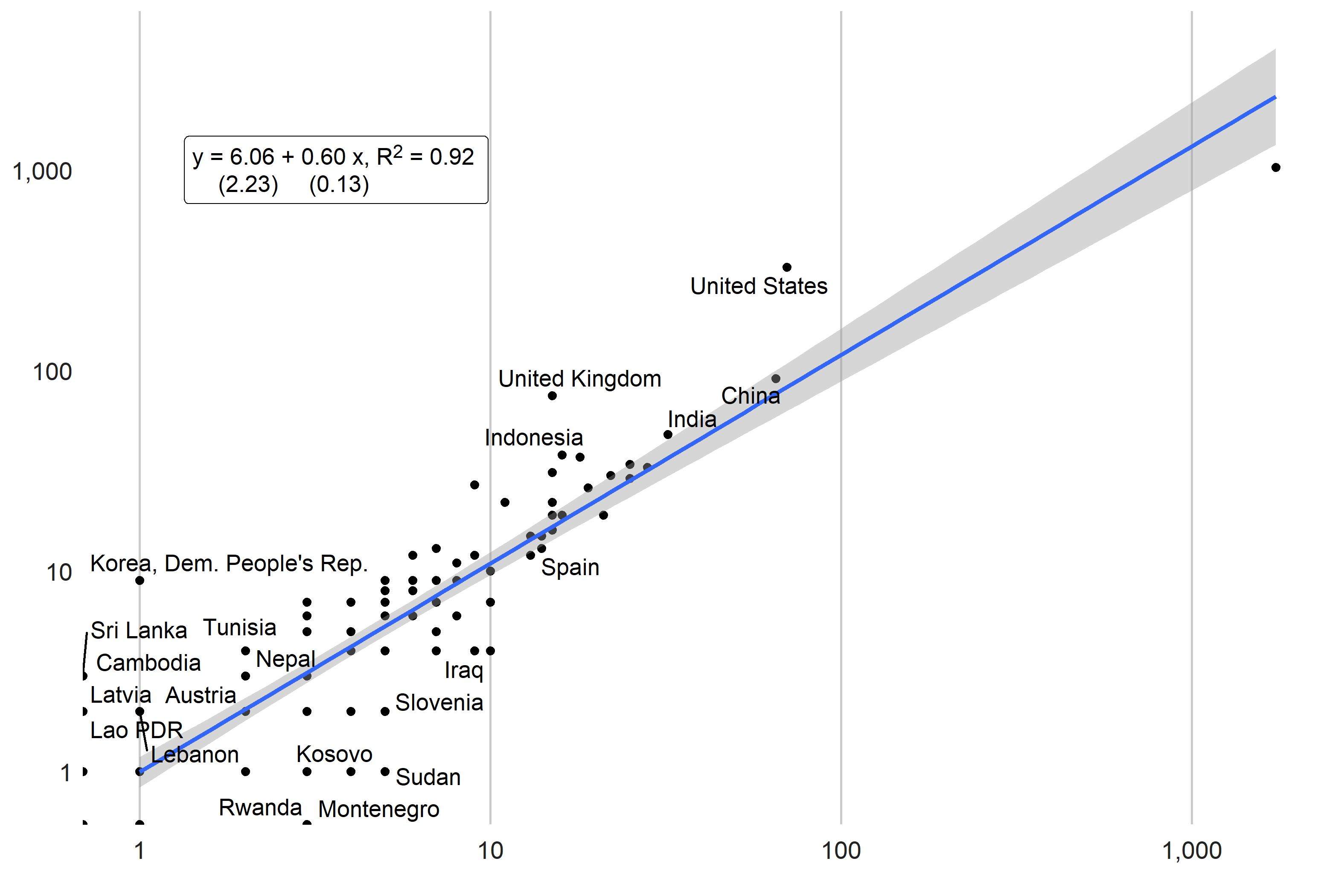
# Analysis

## Comparison of Mturk Ratings of Articles to Machine Learning Approach

In what follows, a comparison is given between the human raters of the articles, who were hired through Amazon Mechanical Turk (MTurk), and the predictions of the country of study and whether the article uses data from the machine learning model. To do so, a set of 3,500 articles are examined that were hand coded by the MTurk workers. This set of 3,500 articles is then fed to the machine learning model, from which the country of study and data use status are identified.

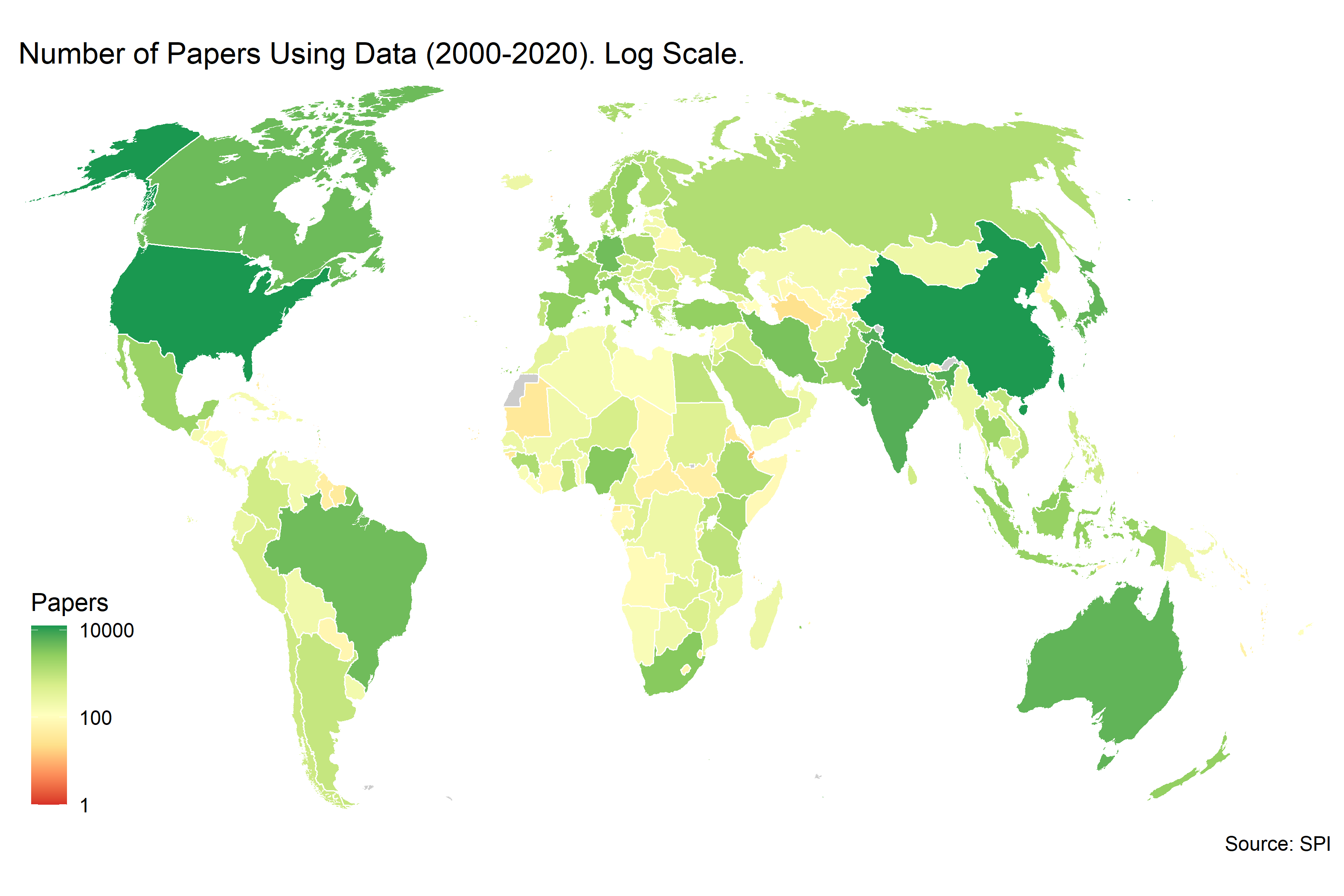
Among this set of 3,500 articles, we can see the correlation between the number of articles written about each country using the two approaches and percentage of articles classified as using data. The Pearson between the human raters and the NLP predictions is 0.96. The figure below shows the scatterplot between the human classifications and the NLP classifications.

While there is strong agreement between the two approaches, there are some notable differences. First, some countries are notably over-represented among human classifiers compared to those produced using the machine learning model. One is North Korea, which was selected by human raters nine times, while the NLP model selected only 1 paper as related to North Korea. Upon inspecting the nine articles identified by the Amazon MTurk classifiers, it appears that these articles were misclassified and were actually referring to South Korea. Additionally, the United Kingdom and United States are notable outliers. It is possible other contextual information in the title or abstract of the article, such as mention of a specific policy related to the country, led the human classifiers to classify more articles than could be identified using the regular expressions search for country names and the named entity recognition tools.

Figure. Comparison of Human Classifications of Data Use to NLP Predictions 

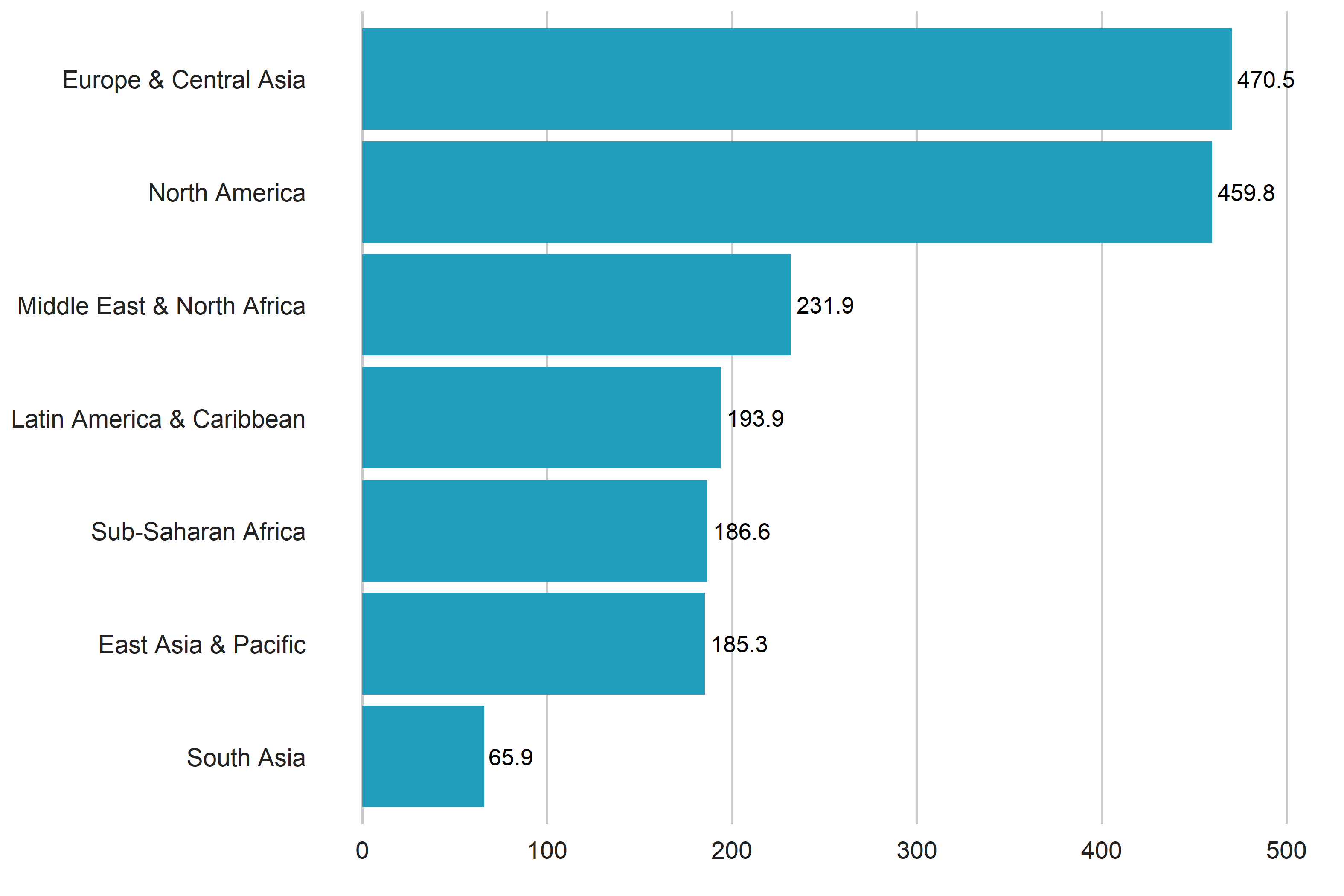
## Results

Using the NLP model results, the number of articles using data produced for each country is shown in the figure below. Around 140,025 articles could be identified with a particular country. The two countries with the largest number of papers using data produced are the United States (12,273 papers) and China (12,063) papers between 2000 and 2020. India, Australia, and Japan are third, fourth, and fifth respectively with 6,481, 5,463, and 5,300 papers respectively.

Figure. Number of Articles using Data by Country (2000-2020) 

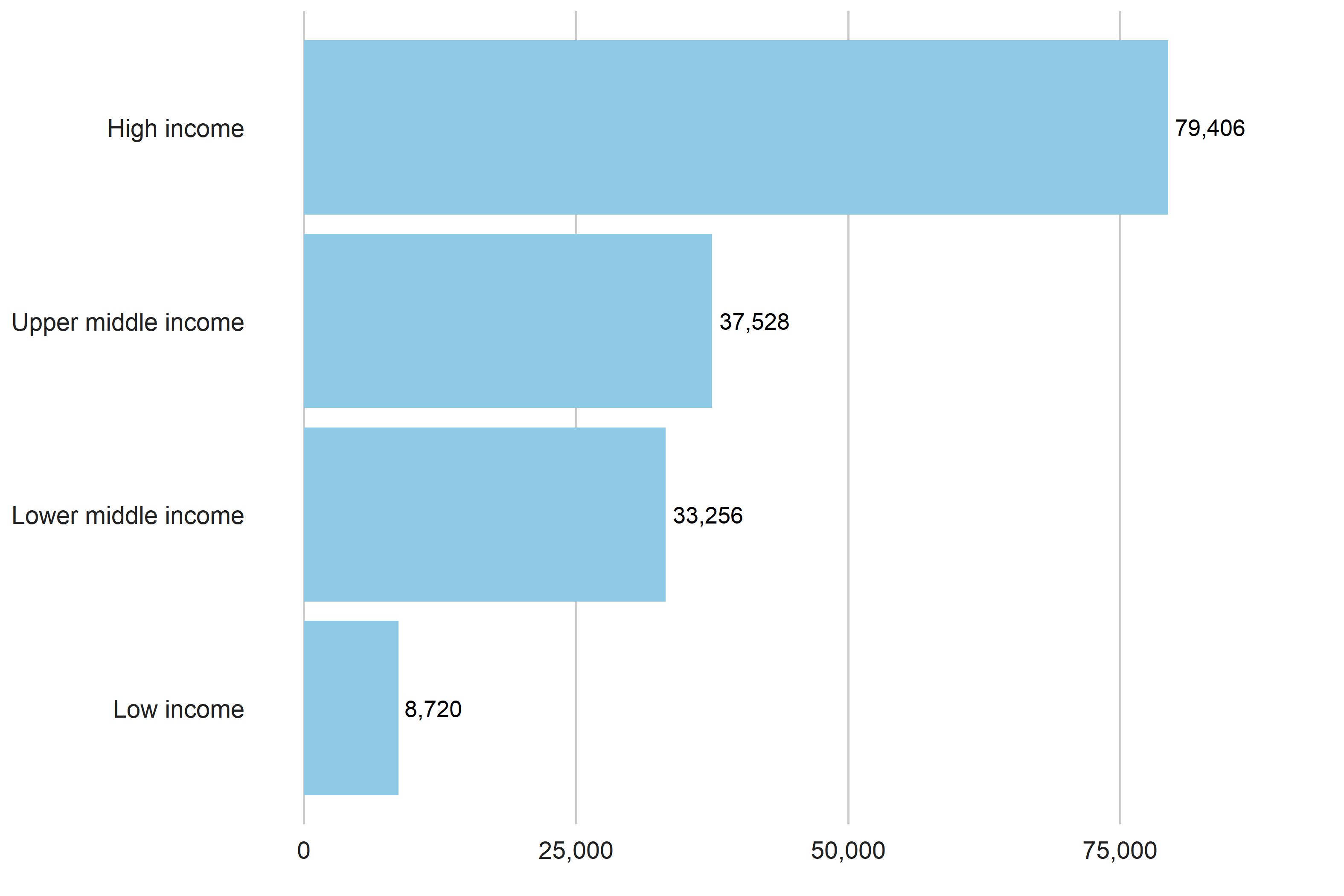
The numbers in the figure below shows the number of articles using data produced per million people in each region. When looking by region, Europe and Central Asia had the largest number of articles using data produced on a per capita basis with around 470 articles produced per million. North America is second with around 460 per million. South Asia produced the fewest number of papers per million with only around 65 articles produced per million persons.

Figure. Number of articles using data by region per million persons

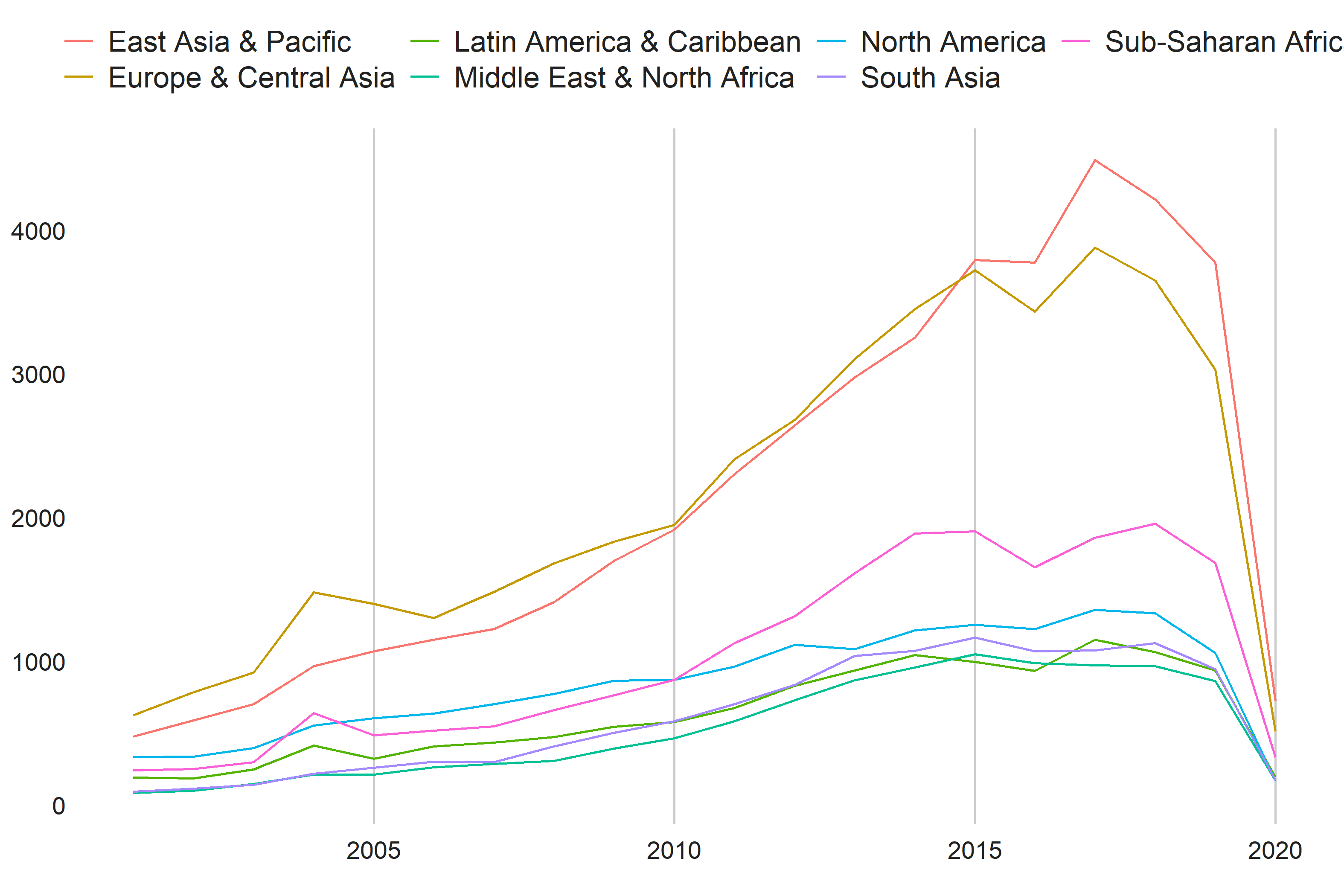


Among income groups, high income countries produced nearly 50% of all papers using data from 2000-2020, despite only making up around 17% of the world’s population. Despite making up around 1/3rd of the world population, low income countries only produce around 5% of articles using data.

Figure. TOtal Number of articles using data by income



The chart below shows the number of articles using data by region over time. Note that only papers released prior to April 2020 are included in the S2ORC dataset currently, so in the year 2020 there is a reduced number of papers released.

Figure. Articles Using Data over Time. 

There is a strong relationship between the number of articles produced using data and GDP, Population, and the SPI Overall score of a country. Bivariate Regression coefficients from a regression of log number of data use papers on GDP, Population, and the SPI overall score is shown within the figures. In the case of GDP and population, the regressions are on logged values of GDP or population as the case may be.

For GDP, the figures indicate an elasticity of 0.6, meaning that for instance a 10% increase in GDP translate into a 6% increase in academic articles using data. Log GDP explains 77% of the variation () in the number of articles produced using data. The elasticity with respect to population is similar in magnitude to that of GDP, and population alone explains 69% of the variation in academic articles using data.

SPI overall scores, a measure of statistical performance, also are strongly predictive of academic output using data. A 10 point increase in SPI scores, which is approximately the same as moving from the median SPI score to the 65th percentile, translates into around a .5% increase in the number of articles using data.

Figure. Relationship between Papers using Data and Development Outcomes

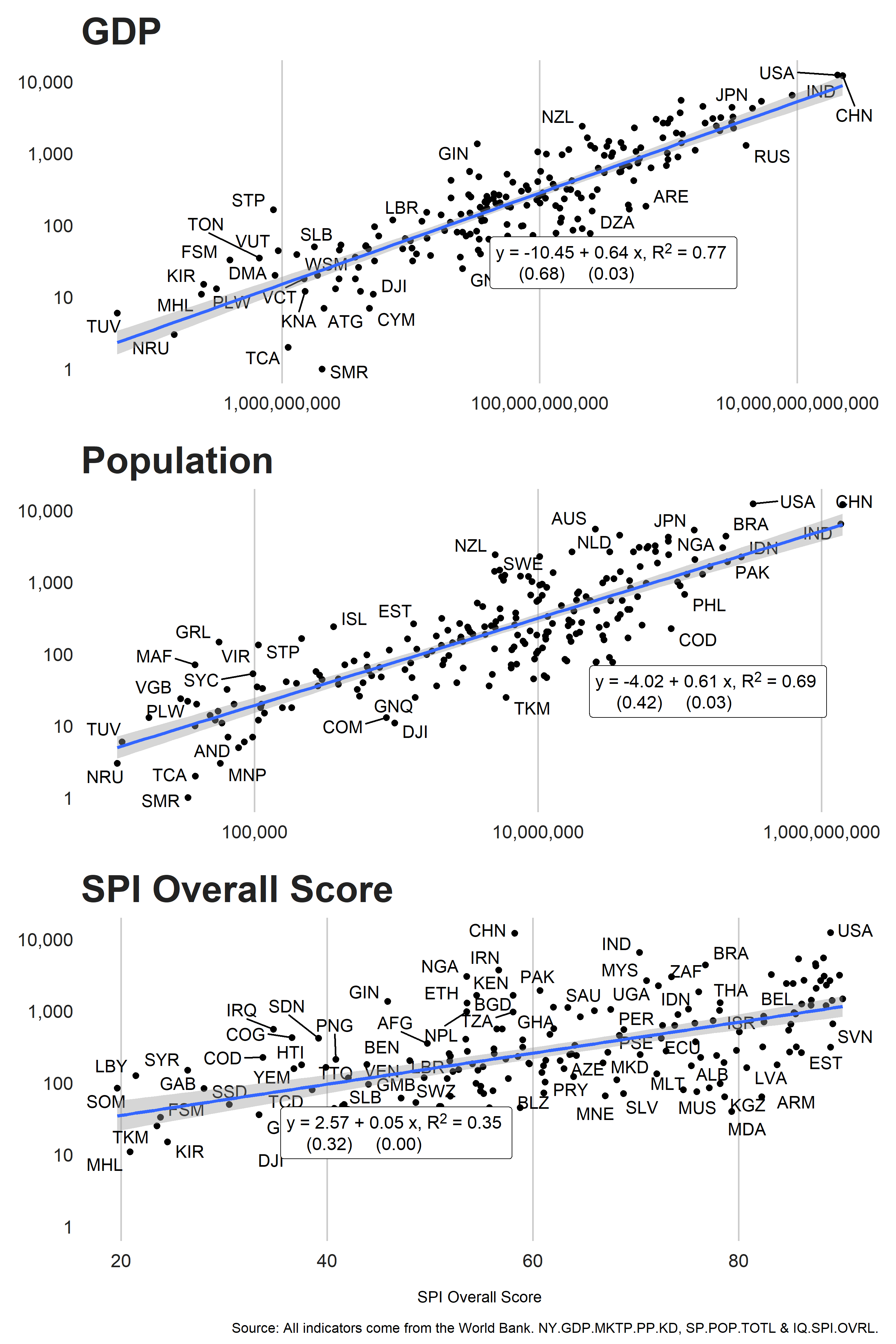
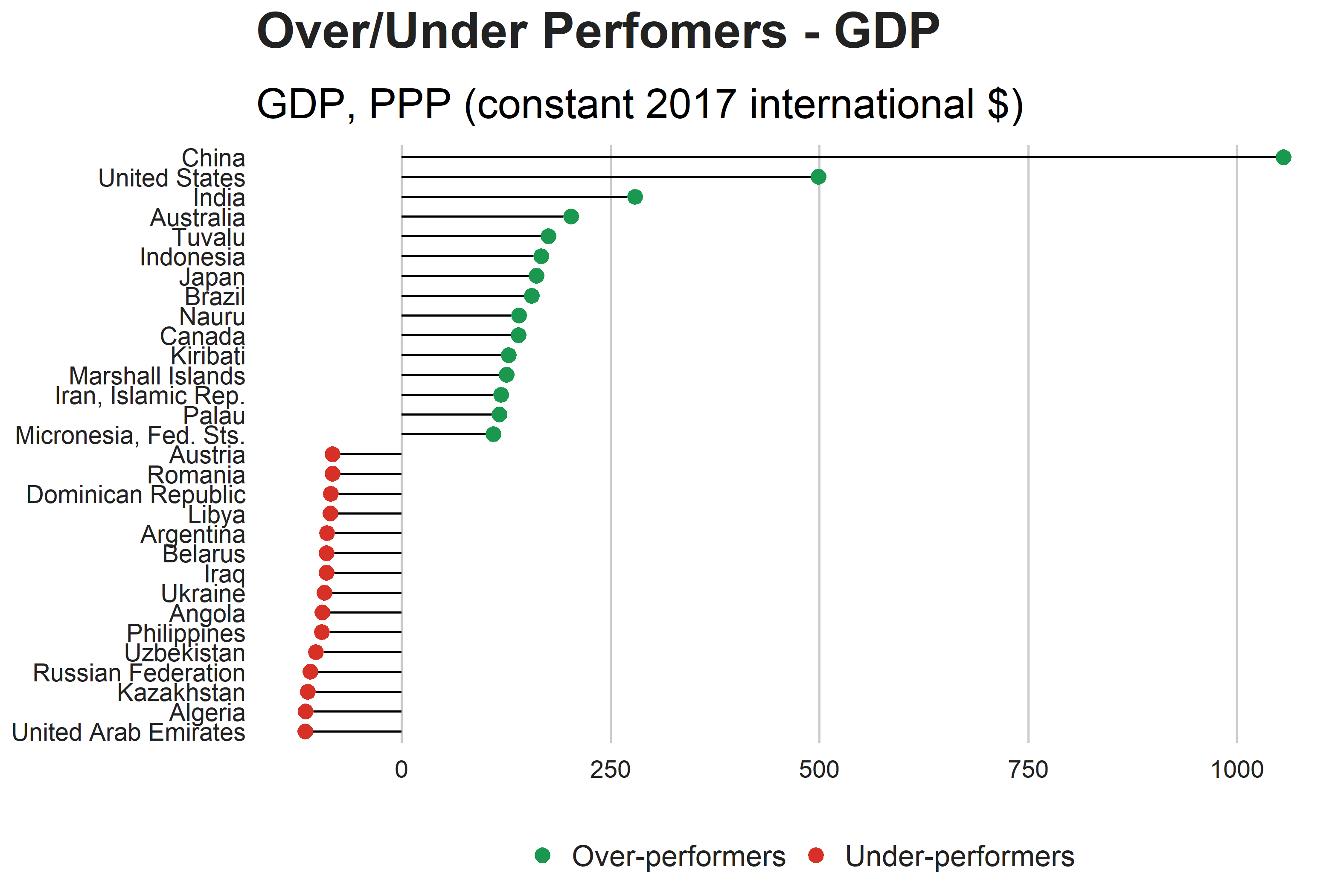


Table. Relationships between Number of Papers Using Data and Statistical Performance Scores

|  | GDP | Population | GDP + Population | SPI | SPI + Population + GDP |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | -9.38\*\*\* | -4.16\*\*\* | -9.01\*\*\* | 2.48\*\*\* | -7.50\*\*\* |
|  | (0.68) | (0.50) | (0.71) | (0.34) | (0.84) |
| Log GDP | 0.60\*\*\* |  | 0.49\*\*\* |  | 0.30\*\*\* |
|  | (0.03) |  | (0.05) |  | (0.07) |
| Log Population |  | 0.62\*\*\* | 0.14\* |  | 0.27\*\*\* |
|  |  | (0.03) | (0.06) |  | (0.06) |
| SPI Overall Score |  |  |  | 0.05\*\*\* | 0.02\*\*\* |
|  |  |  |  | (0.01) | (0.01) |
| N | 170 | 170 | 170 | 170 | 170 |
| R Sq. | 0.73 | 0.60 | 0.74 | 0.34 | 0.76 |
|  | | | | | |

Below we show the countries that over or underperform GDP, PPP in 2019.

##   
## Call:  
## lm(formula = data\_papers ~ log(NY.GDP.MKTP.PP.KD), data = gdp\_df\_2019)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -115.69 -55.37 -24.33 29.27 1055.65   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -758.303 81.435 -9.312 <2e-16 \*\*\*  
## log(NY.GDP.MKTP.PP.KD) 32.873 3.247 10.124 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 107.7 on 191 degrees of freedom  
## (24 observations deleted due to missingness)  
## Multiple R-squared: 0.3492, Adjusted R-squared: 0.3458   
## F-statistic: 102.5 on 1 and 191 DF, p-value: < 2.2e-16



## Comparison to Das et al. (2013) and Porteous (2022)

Figure. Comparison to Number of Academic Articles in Das, Do, Shaines, and Srikant (2013)

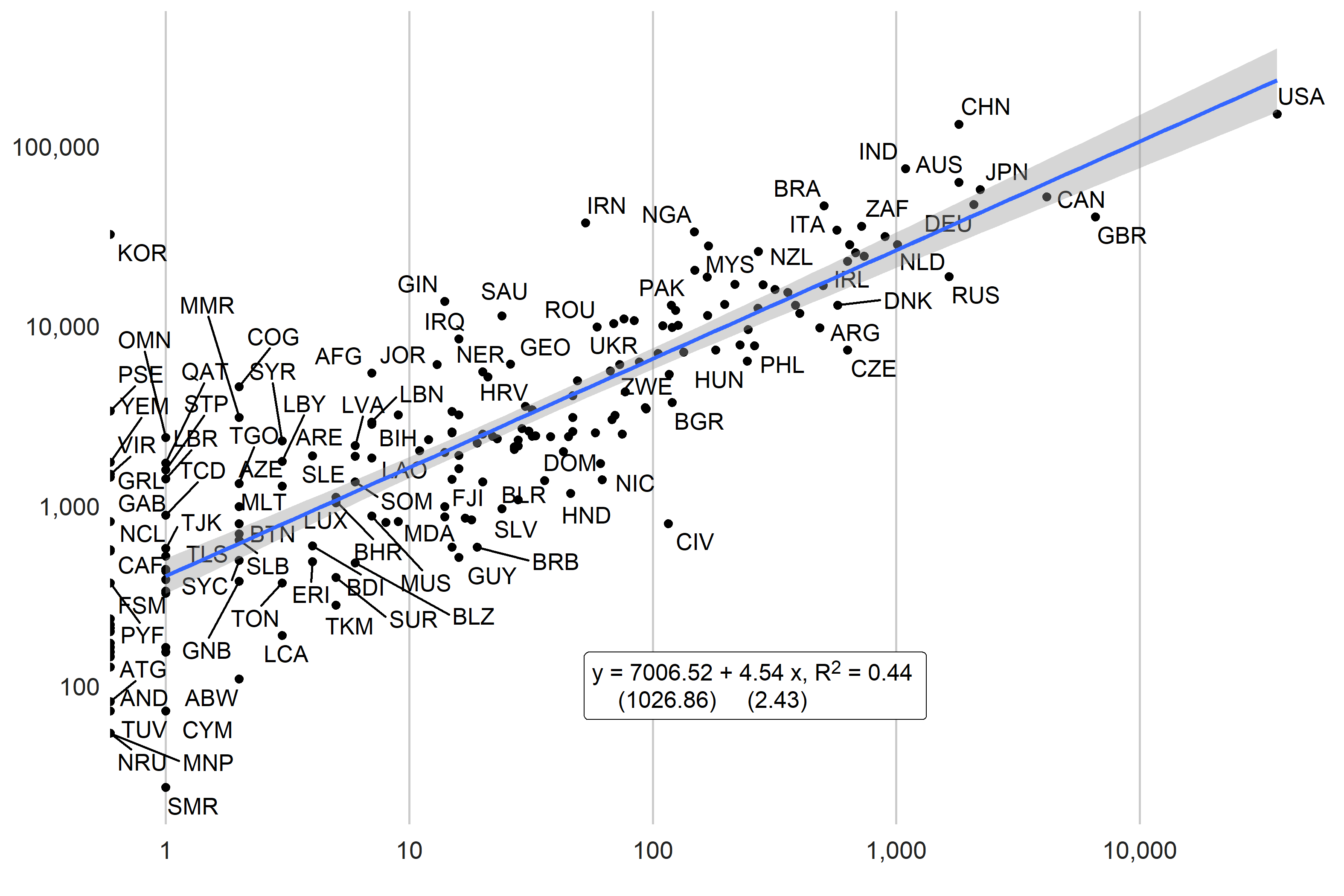
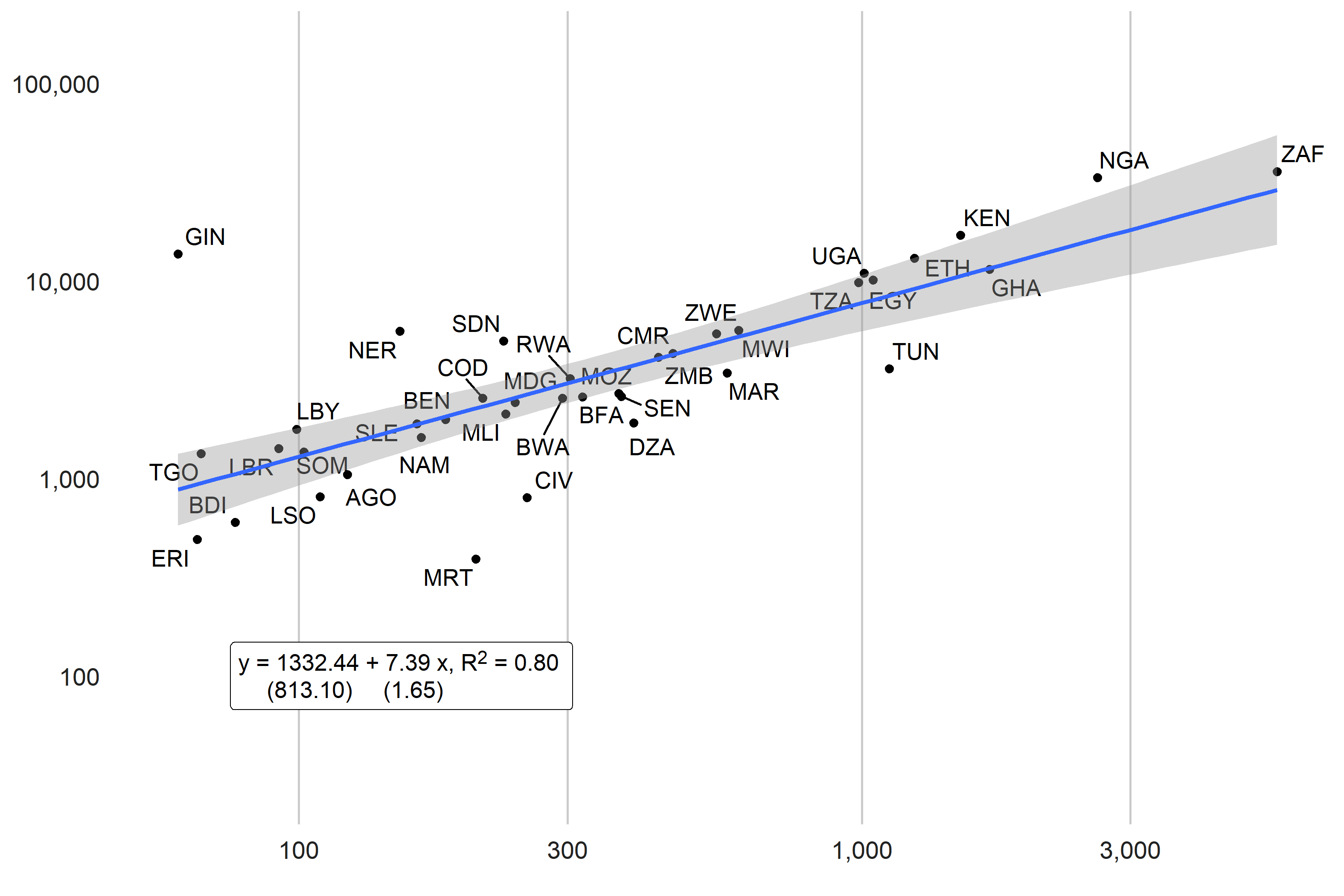


Figure. Comparison to Number of Academic Articles in Porteous (2022) 

## Country Rankings

The S2ORC database contains papers released up to April 14, 2020. Because the year 2020 is only partially captured, annual country scores on academic data use will be released up until 2019. Future data collection will take place to incorporate papers after April 2020.

Three measures of data use by academia will be constructed as follows. These measures will be produced for countries on an annual basis, and a time series will be created based on the year of the publication of the article dating back to year 2002, which is the start of the s2orc dataset time series. Each of the three measures provides slightly different information about a country’s academic data use.

The first use of data measure is a per capita measure, , where for each country i, is the population size of the country and is the number of articles making use of data.

A second measure will be considered including a regression adjusted measure that adjust for other factors in the country such as the population size and GDP This approach responds to the findings in Das et al. (2013), who note in their study of cross-country differences in economic academic output that around 75% of the cross-country variation in economics output is accounted for by population size and income alone. A regression adjusted measure then may be better able to isolate differences in national statistical systems, as opposed to differences between countries in population and income. In this case the use of data measure for country i will be the difference between the actual number of academic articles using data and those predicted based on population and GDP.

where

And and are estimated using regression.

A third measure we will consider is the share of papers on a country that use data. It will be calculated as the ratio of papers that use data to the number of total papers published for the country. A ratio of 1 indicates that all of the papers published for a country used data, while a ratio of zero indicates none of the papers published for a country use data.

Because of volatility in the number of publications for a country in a particular year, especially for countries with relatively few publications, a 3 year moving average will be taken of the number of publications using data for each country.

The table below is sorted based on the regression adjusted measure of data use.

Countries shaded in dark orange are the lowest performing, countries in dark green are the highest performing. Countries are grouped into five groups:

* **Top Quintile**: Countries in the Top quintile are classified in this group. Shading in dark green.
* **4th Quintile**: Countries in the 4th quintile, or those above the 60th percentile but below the 80th percentile are in this group. Shading in light green.
* **3rd Quintile**: Countries in the 3rd quintile, or those between the 40th and 60th percentile, are classified in this group. Shading in yellow.
* **2nd Quintile**: Countries in the 2nd quintile, or those above the 20th percentile but below the 40th percentile, are in this group. Shading in light orange.
* **Bottom 20%**: Countries in the bottom 20% are classified in this group. Shading in dark orange .

Table. Country Scores on Academic Data Use. Year 2019.

| iso3c | Country | Academic Data Use per capita | Academic Data Use Regression Adjusted | Academic Data Use as Fraction of All Papers | Total Academic Publications using Data |
| --- | --- | --- | --- | --- | --- |
| NZL | New Zealand | 37.29 | 2.38 | 0.87 | 140 |
| AUS | Australia | 17.79 | 2.34 | 0.88 | 360 |
| GRL | Greenland | 213.43 | 2.21 | 0.90 | 12 |
| NOR | Norway | 24.56 | 1.99 | 0.89 | 104 |
| SWE | Sweden | 18.13 | 1.97 | 0.88 | 135 |
| CAN | Canada | 9.59 | 1.89 | 0.84 | 310 |
| IRL | Ireland | 22.56 | 1.87 | 0.81 | 79 |
| DNK | Denmark | 19.38 | 1.79 | 0.91 | 96 |
| NLD | Netherlands | 12.05 | 1.78 | 0.88 | 195 |
| ISL | Iceland | 59.17 | 1.72 | 0.85 | 18 |
| FIN | Finland | 17.21 | 1.65 | 0.89 | 74 |
| SGP | Singapore | 16.54 | 1.63 | 0.90 | 89 |
| VIR | Virgin Islands (U.S.) | 90.62 | 1.62 | 0.85 | 6 |
| MYS | Malaysia | 7.44 | 1.56 | 0.89 | 205 |
| USA | United States | 2.70 | 1.55 | 0.81 | 748 |
| CHE | Switzerland | 12.09 | 1.49 | 0.87 | 91 |
| MAF | St. Martin (French part) | 122.80 | 1.49 | 0.82 | 4 |
| KOR | Korea, Rep. | 5.51 | 1.47 | 0.92 | 248 |
| HKG | Hong Kong SAR, China | 11.23 | 1.36 | 0.89 | 80 |
| JPN | Japan | 3.37 | 1.36 | 0.90 | 366 |
| DEU | Germany | 4.01 | 1.35 | 0.87 | 286 |
| IRN | Iran, Islamic Rep. | 3.98 | 1.35 | 0.94 | 276 |
| ZAF | South Africa | 4.58 | 1.34 | 0.83 | 223 |
| STP | Sao Tome and Principe | 49.60 | 1.32 | 0.94 | 11 |
| ITA | Italy | 4.40 | 1.31 | 0.89 | 245 |
| ESP | Spain | 4.79 | 1.29 | 0.89 | 214 |
| GBR | United Kingdom | 3.83 | 1.21 | 0.78 | 222 |
| CHN | China | 0.99 | 1.17 | 0.90 | 1,308 |
| GRC | Greece | 7.37 | 1.09 | 0.78 | 72 |
| ISR | Israel | 7.55 | 1.04 | 0.78 | 52 |
| PRT | Portugal | 7.00 | 1.02 | 0.80 | 61 |
| GIB | Gibraltar | 79.12 | 1.00 | 0.80 | 3 |
| BRA | Brazil | 1.84 | 0.98 | 0.90 | 343 |
| GEO | Georgia | 10.39 | 0.98 | 0.84 | 36 |
| SAU | Saudi Arabia | 3.84 | 0.93 | 0.91 | 134 |
| FRA | France | 2.85 | 0.92 | 0.83 | 168 |
| MCO | Monaco | 68.43 | 0.91 | 1.00 | 2 |
| EST | Estonia | 14.82 | 0.89 | 0.77 | 17 |
| AUT | Austria | 6.49 | 0.88 | 0.85 | 44 |
| GHA | Ghana | 3.85 | 0.88 | 0.93 | 111 |
| BEL | Belgium | 5.74 | 0.87 | 0.85 | 65 |
| KEN | Kenya | 2.94 | 0.85 | 0.91 | 136 |
| POL | Poland | 3.37 | 0.85 | 0.85 | 138 |
| TUR | Turkiye | 2.33 | 0.81 | 0.87 | 169 |
| HRV | Croatia | 8.36 | 0.80 | 0.84 | 30 |
| BRN | Brunei Darussalam | 20.77 | 0.75 | 0.96 | 8 |
| SVN | Slovenia | 10.53 | 0.74 | 0.75 | 19 |
| COG | Congo, Rep. | 6.88 | 0.72 | 0.86 | 32 |
| IDN | Indonesia | 1.28 | 0.71 | 0.81 | 355 |
| CZE | Czech Republic | 4.97 | 0.69 | 0.87 | 45 |
| NPL | Nepal | 3.22 | 0.68 | 0.92 | 69 |
| BWA | Botswana | 9.26 | 0.66 | 0.83 | 18 |
| PLW | Palau | 74.07 | 0.66 | 1.00 | 1 |
| JOR | Jordan | 4.88 | 0.65 | 0.87 | 54 |
| THA | Thailand | 2.13 | 0.65 | 0.94 | 146 |
| GIN | Guinea | 4.38 | 0.64 | 0.92 | 55 |
| NGA | Nigeria | 1.35 | 0.64 | 0.86 | 248 |
| CHL | Chile | 3.57 | 0.61 | 0.89 | 61 |
| CYP | Cyprus | 11.12 | 0.56 | 0.77 | 9 |
| LTU | Lithuania | 7.64 | 0.55 | 0.77 | 15 |
| ETH | Ethiopia | 1.55 | 0.53 | 0.93 | 187 |
| ROU | Romania | 3.27 | 0.53 | 0.79 | 49 |
| FJI | Fiji | 11.99 | 0.51 | 0.70 | 10 |
| SYC | Seychelles | 30.73 | 0.51 | 1.00 | 2 |
| UGA | Uganda | 2.23 | 0.50 | 0.93 | 88 |
| SVK | Slovak Republic | 5.44 | 0.49 | 0.83 | 19 |
| QAT | Qatar | 7.06 | 0.48 | 0.86 | 23 |
| MWI | Malawi | 3.08 | 0.45 | 0.95 | 50 |
| LVA | Latvia | 8.01 | 0.43 | 0.78 | 13 |
| MNG | Mongolia | 6.41 | 0.43 | 0.97 | 16 |
| PRI | Puerto Rico | 6.26 | 0.41 | 0.92 | 19 |
| LKA | Sri Lanka | 2.71 | 0.39 | 0.87 | 47 |
| MEX | Mexico | 1.25 | 0.37 | 0.85 | 143 |
| VGB | British Virgin Islands | 44.40 | 0.37 | 1.00 | 1 |
| HUN | Hungary | 3.68 | 0.35 | 0.79 | 31 |
| BMU | Bermuda | 31.29 | 0.34 | 0.86 | 0 |
| SRB | Serbia | 4.22 | 0.34 | 0.77 | 21 |
| OMN | Oman | 4.69 | 0.31 | 0.92 | 17 |
| BTN | Bhutan | 10.05 | 0.27 | 0.92 | 5 |
| KNA | St. Kitts and Nevis | 31.55 | 0.27 | 1.00 | 1 |
| MLT | Malta | 11.90 | 0.26 | 0.95 | 7 |
| LUX | Luxembourg | 10.75 | 0.25 | 0.74 | 4 |
| TZA | Tanzania | 1.54 | 0.24 | 0.91 | 77 |
| IND | India | 0.39 | 0.23 | 0.82 | 502 |
| TUV | Tuvalu | 57.20 | 0.22 | 1.00 | 2 |
| GUM | Guam | 17.93 | 0.20 | 0.75 | 1 |
| BGD | Bangladesh | 0.94 | 0.19 | 0.91 | 121 |
| KWT | Kuwait | 4.44 | 0.18 | 0.98 | 16 |
| PAK | Pakistan | 0.81 | 0.16 | 0.89 | 180 |
| BLZ | Belize | 11.96 | 0.15 | 0.93 | 5 |
| KIR | Kiribati | 19.84 | 0.15 | 1.00 | 5 |
| VUT | Vanuatu | 13.34 | 0.15 | 1.00 | 7 |
| MNE | Montenegro | 9.65 | 0.14 | 0.90 | 5 |
| PSE | West Bank and Gaza | 3.98 | 0.12 | 0.77 | 19 |
| ZWE | Zimbabwe | 2.46 | 0.12 | 0.81 | 31 |
| MKD | North Macedonia | 5.62 | 0.11 | 0.74 | 8 |
| VNM | Vietnam | 1.08 | 0.11 | 0.90 | 104 |
| XKX | Kosovo | 5.96 | 0.11 | 0.63 | 9 |
| FRO | Faroe Islands | 27.39 | 0.09 | 1.00 | 4 |
| BGR | Bulgaria | 3.25 | 0.08 | 0.81 | 22 |
| COL | Colombia | 1.38 | 0.08 | 0.84 | 72 |
| LBN | Lebanon | 3.21 | 0.06 | 0.87 | 21 |
| MAC | Macao SAR, China | 8.85 | 0.06 | 0.89 | 2 |
| ABW | Aruba | 18.81 | 0.05 | 1.00 | 0 |
| CRI | Costa Rica | 3.57 | 0.04 | 0.90 | 16 |
| NAM | Namibia | 4.81 | 0.04 | 0.78 | 13 |
| ASM | American Samoa | 24.11 | 0.02 | 0.80 | 2 |
| GRD | Grenada | 17.86 | 0.02 | 0.86 | 2 |
| FSM | Micronesia, Fed. Sts. | 17.57 | 0.01 | 0.75 | 2 |
| BHS | Bahamas, The | 10.27 | 0.00 | 1.00 | 2 |
| PER | Peru | 1.55 | 0.00 | 0.83 | 43 |
| ARG | Argentina | 1.31 | -0.03 | 0.81 | 60 |
| LIE | Liechtenstein | 26.30 | -0.05 | 0.75 | 0 |
| TUN | Tunisia | 2.28 | -0.05 | 0.87 | 24 |
| RUS | Russian Federation | 0.77 | -0.06 | 0.67 | 86 |
| BIH | Bosnia and Herzegovina | 3.84 | -0.07 | 0.69 | 11 |
| URY | Uruguay | 3.76 | -0.07 | 0.83 | 12 |
| NER | Niger | 1.63 | -0.09 | 0.93 | 39 |
| ZMB | Zambia | 1.83 | -0.09 | 0.90 | 33 |
| PYF | French Polynesia | 10.74 | -0.10 | 0.82 | 4 |
| BRB | Barbados | 10.45 | -0.11 | 0.82 | 1 |
| LAO | Lao PDR | 2.65 | -0.11 | 0.92 | 20 |
| RWA | Rwanda | 2.06 | -0.12 | 0.87 | 27 |
| TON | Tonga | 15.95 | -0.12 | 1.00 | 2 |
| ECU | Ecuador | 1.77 | -0.14 | 0.84 | 21 |
| JAM | Jamaica | 3.73 | -0.15 | 0.83 | 5 |
| SLB | Solomon Islands | 6.97 | -0.15 | 0.88 | 3 |
| VCT | St. Vincent and the Grenadines | 15.07 | -0.15 | 1.00 | 2 |
| WSM | Samoa | 11.84 | -0.15 | 0.70 | 3 |
| SWZ | Eswatini | 5.52 | -0.16 | 0.90 | 5 |
| CMR | Cameroon | 1.44 | -0.17 | 0.90 | 43 |
| PAN | Panama | 3.06 | -0.19 | 0.89 | 11 |
| BHR | Bahrain | 4.47 | -0.22 | 0.85 | 6 |
| EGY | Egypt, Arab Rep. | 0.76 | -0.22 | 0.86 | 83 |
| IMN | Isle of Man | 15.76 | -0.22 | 1.00 | 1 |
| SUR | Suriname | 6.88 | -0.23 | 1.00 | 4 |
| ALB | Albania | 3.39 | -0.26 | 0.72 | 6 |
| KHM | Cambodia | 1.52 | -0.31 | 0.85 | 25 |
| ARE | United Arab Emirates | 1.84 | -0.34 | 0.87 | 21 |
| UKR | Ukraine | 0.96 | -0.34 | 0.59 | 37 |
| MDV | Maldives | 6.28 | -0.36 | 0.91 | 5 |
| SLE | Sierra Leone | 1.96 | -0.37 | 0.79 | 15 |
| BFA | Burkina Faso | 1.28 | -0.39 | 0.99 | 24 |
| SEN | Senegal | 1.37 | -0.42 | 0.88 | 20 |
| NRU | Nauru | 30.97 | -0.43 | 1.00 | 0 |
| IRQ | Iraq | 0.92 | -0.44 | 0.75 | 32 |
| TCA | Turks and Caicos Islands | 17.45 | -0.46 | 1.00 | 1 |
| NCL | New Caledonia | 7.37 | -0.49 | 1.00 | 1 |
| GMB | Gambia, The | 2.84 | -0.52 | 0.95 | 4 |
| PHL | Philippines | 0.55 | -0.52 | 0.81 | 53 |
| LBR | Liberia | 2.03 | -0.54 | 0.94 | 9 |
| SDN | Sudan | 0.80 | -0.54 | 0.87 | 30 |
| BEN | Benin | 1.36 | -0.56 | 0.89 | 15 |
| TTO | Trinidad and Tobago | 3.35 | -0.57 | 0.82 | 2 |
| MUS | Mauritius | 3.42 | -0.59 | 0.76 | 3 |
| MAR | Morocco | 0.79 | -0.63 | 0.88 | 28 |
| TGO | Togo | 1.48 | -0.64 | 0.92 | 8 |
| ARM | Armenia | 2.25 | -0.65 | 0.77 | 2 |
| LCA | St. Lucia | 7.29 | -0.66 | 0.80 | 0 |
| PNG | Papua New Guinea | 1.41 | -0.66 | 0.82 | 14 |
| MNP | Northern Mariana Islands | 11.65 | -0.69 | 0.67 | 0 |
| MMR | Myanmar | 0.62 | -0.70 | 0.78 | 40 |
| MHL | Marshall Islands | 11.34 | -0.71 | 1.00 | 1 |
| NIC | Nicaragua | 1.48 | -0.73 | 0.91 | 11 |
| CPV | Cabo Verde | 4.24 | -0.74 | 1.00 | 2 |
| MLI | Mali | 0.92 | -0.74 | 0.84 | 20 |
| BOL | Bolivia | 1.13 | -0.76 | 0.70 | 13 |
| CYM | Cayman Islands | 10.26 | -0.77 | 0.67 | 1 |
| MOZ | Mozambique | 0.74 | -0.77 | 0.89 | 21 |
| CUB | Cuba | 1.09 | -0.80 | 0.70 | 12 |
| HTI | Haiti | 1.10 | -0.80 | 0.82 | 11 |
| DMA | Dominica | 9.28 | -0.82 | 1.00 | 1 |
| DOM | Dominican Republic | 1.06 | -0.86 | 0.83 | 12 |
| MDG | Madagascar | 0.68 | -0.90 | 0.96 | 15 |
| KAZ | Kazakhstan | 0.79 | -0.91 | 0.73 | 14 |
| SYR | Syrian Arab Republic | 0.80 | -0.93 | 0.62 | 16 |
| GUY | Guyana | 2.98 | -0.94 | 1.00 | 2 |
| AFG | Afghanistan | 0.53 | -0.99 | 0.72 | 15 |
| ATG | Antigua and Barbuda | 6.86 | -1.00 | 0.67 | 0 |
| LBY | Libya | 1.08 | -1.03 | 0.67 | 2 |
| GAB | Gabon | 1.69 | -1.07 | 1.00 | 6 |
| AZE | Azerbaijan | 0.86 | -1.08 | 0.72 | 9 |
| GTM | Guatemala | 0.70 | -1.08 | 0.70 | 7 |
| SMR | San Marino | 9.84 | -1.09 | 1.00 | 0 |
| TLS | Timor-Leste | 2.06 | -1.09 | 0.80 | 3 |
| HND | Honduras | 0.86 | -1.11 | 0.81 | 9 |
| CAF | Central African Republic | 1.12 | -1.14 | 0.80 | 3 |
| LSO | Lesotho | 1.57 | -1.15 | 1.00 | 4 |
| YEM | Yemen, Rep. | 0.47 | -1.24 | 0.82 | 17 |
| CUW | Curacao | 4.23 | -1.27 | 1.00 | 0 |
| PRY | Paraguay | 0.80 | -1.31 | 0.81 | 9 |
| GNB | Guinea-Bissau | 1.39 | -1.32 | 0.73 | 1 |
| SLV | El Salvador | 0.77 | -1.38 | 0.83 | 4 |
| MDA | Moldova | 1.13 | -1.39 | 0.56 | 3 |
| GNQ | Equatorial Guinea | 1.47 | -1.41 | 1.00 | 3 |
| COD | Congo, Dem. Rep. | 0.25 | -1.42 | 0.86 | 18 |
| DZA | Algeria | 0.33 | -1.42 | 0.90 | 12 |
| VEN | Venezuela, RB | 0.37 | -1.47 | 0.74 | 7 |
| BLR | Belarus | 0.57 | -1.53 | 0.67 | 4 |
| KGZ | Kyrgyz Republic | 0.67 | -1.53 | 0.59 | 4 |
| SOM | Somalia | 0.45 | -1.55 | 0.66 | 7 |
| SSD | South Sudan | 0.51 | -1.57 | 0.68 | 3 |
| DJI | Djibouti | 1.37 | -1.62 | 0.67 | 1 |
| MRT | Mauritania | 0.66 | -1.69 | 0.82 | 3 |
| TJK | Tajikistan | 0.43 | -1.82 | 0.71 | 2 |
| COM | Comoros | 1.18 | -1.83 | 0.75 | 1 |
| TCD | Chad | 0.31 | -1.90 | 0.75 | 4 |
| BDI | Burundi | 0.35 | -1.94 | 0.75 | 4 |
| CIV | Cote d'Ivoire | 0.25 | -1.94 | 0.90 | 3 |
| AGO | Angola | 0.17 | -2.23 | 0.67 | 4 |
| PRK | Korea, Dem. People's Rep. | 0.14 | -2.48 | 0.50 | 2 |
| UZB | Uzbekistan | 0.12 | -2.55 | 0.80 | 1 |
| TKM | Turkmenistan | 0.17 | -2.94 | 0.50 | 0 |
| AND | Andorra | 0.00 | -Inf | 0.00 | 0 |
| SXM | Sint Maarten (Dutch part) | 0.00 | -Inf |  | 0 |
| ERI | Eritrea |  |  | 0.86 | 5 |
| TWN | Taiwan, China |  |  | 0.90 | 146 |

## Conclusions

# References

Cameron, Drew B, Anjini Mishra, and Annette N Brown. 2016. “The Growth of Impact Evaluation for International Development: How Much Have We Learned?” *Journal of Development Effectiveness* 8 (1): 1–21.

Dang, Hai-Anh H, John Pullinger, Umar Serajuddin, and Brian Stacy. 2021. “Statistical Performance Indicators and Index.”

Das, Jishnu, Quy-Toan Do, Karen Shaines, and Sowmya Srikant. 2013. “US and Them: The Geography of Academic Research.” *Journal of Development Economics* 105: 112–30.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. “Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach.” *The Quarterly Journal of Economics* 133 (2): 801–70.

Kleinberg, Bennett, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. 2018. “Using Named Entities for Computer-Automated Verbal Deception Detection.” *Journal of Forensic Sciences* 63 (3): 714–23.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. “S2ORC: The Semantic Scholar Open Research Corpus.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–83. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.447>.

Porteous, Obie. 2020. “Research Deserts and Oases: Evidence from 27 Thousand Economics Journal Articles on Africa.” *Oxford Bulletin of Economics and Statistics*.

Robinson, Michael D, James E Hartley, and Patricia Higino Schneider. 2006. “Which Countries Are Studied Most by Economists? An Examination of the Regional Distribution of Economic Research.” *Kyklos* 59 (4): 611–26.

Sabet, Shayda Mae, and Annette N Brown. 2018. “Is Impact Evaluation Still on the Rise? The New Trends in 2010–2015.” *Journal of Development Effectiveness* 10 (3): 291–304.

Shelar, Hemlata, Gagandeep Kaur, Neha Heda, and Poorva Agrawal. 2020. “Named Entity Recognition Approaches and Their Comparison for Custom Ner Model.” *Science & Technology Libraries* 39 (3): 324–37.

Yu Tian. 2020. “THE PARTNER REPORT ON SUPPORT TO STATISTICS PRESS 2020.” PARIS21. <https://paris21.org/sites/default/files/inline-files/PRESS2020%20Final.pdf>.

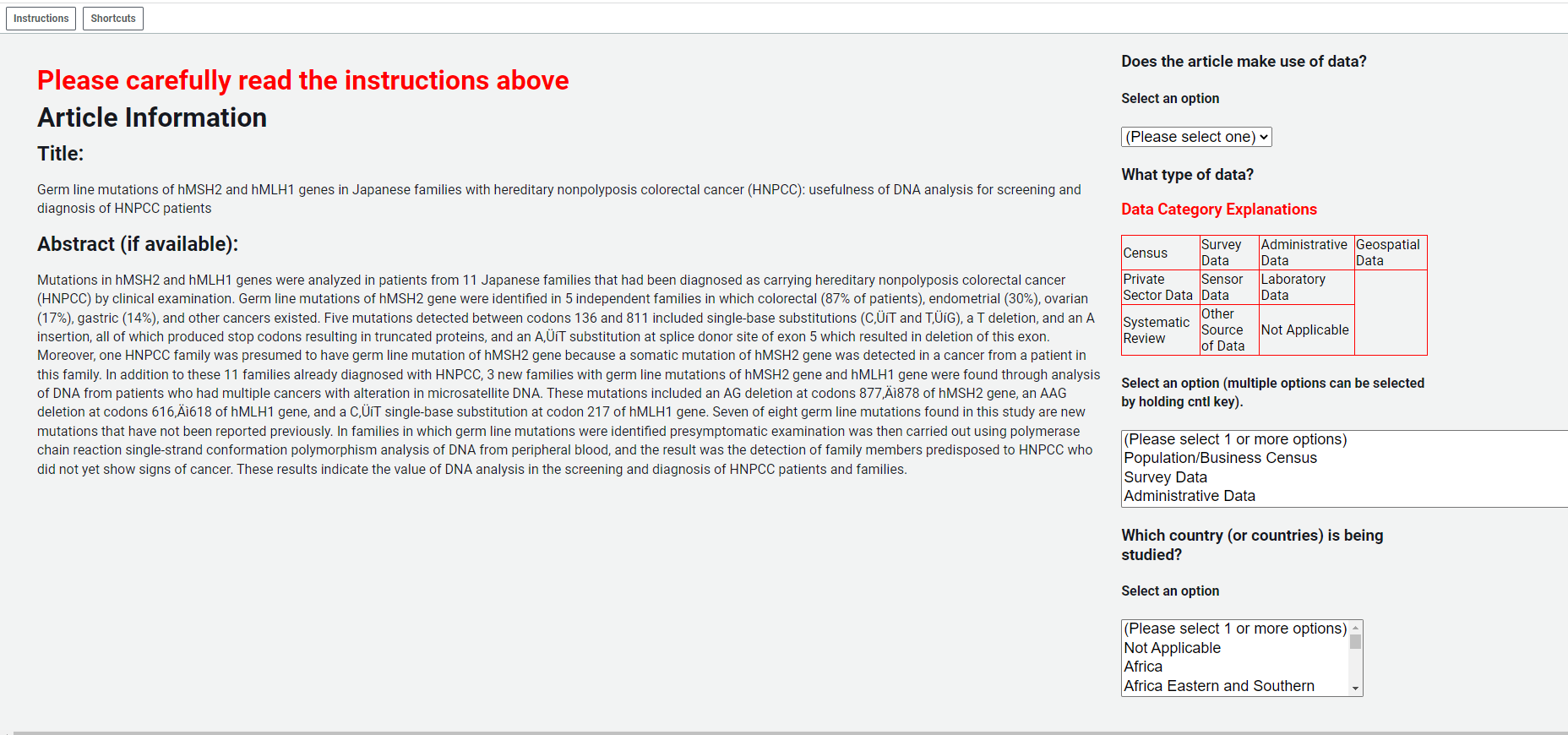
# Appendix

## Amazon Mturk Prompt

Amazon MTurk workers were given the following instructions before completing the activity.

Each of these documents is an academic article. The goal of this study is to measure whether a specific academic article is using data and from which country the data came.  
  
There are two classification tasks in this exercise:  
(1) identifying whether an academic article is using data from a country  
(2) identifying from which country that data came.  
  
For task 1, we are looking specifically at the use of data. Data is any information that has been collected, observed, generated or created to produce research findings. As an example, a study that reports findings or analysis using a survey data, uses data. Some clues to indicate that a study does use data includes whether a survey or census is described, a statistical model estimated, or a table or means or summary statistics is reported.  
  
After an article is classified as using data, please note the type of data used. The options are population or business census, survey data, administrative data, geospatial data, private sector data, and other data. If no data is used, then mark "Not applicable". In cases where multiple data types are used, please hold the cntl key and click multiple options.  
  
For task 2, we are looking at the country or countries that are studied in the article. In some cases, no country may be applicable. For instance, if the research is theoretical and has no specific country application. In some cases, the research article may involve multiple countries. In these cases, please hold down the cntl key and select all countries that are discussed in the paper.  
  
We expect between 10 and 35 percent of all articles to use data.

Below is an image of an example article for classification.



1. The World Development Indicators includes 217 countries. The country not found with a paper in the WDI is Sint Maarten (Dutch part). [↑](#footnote-ref-20)