Where are the Data Deserts? Measuring the Use of Data by Academia Across Countries using Natural Language Processing

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

2022-11-22

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 0.99 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.

# 1 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 218 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, this paper proposes a new methodology for measuring data use by academia. The measures produced are then used to study the relationship between products produced by the national statistical system and research output.

After training a natural language processing model on around 900 hand coded articles using Amazon Mturk workers, the model was able to predict whether an article made use of data with 87% accuracy compared to a set of articles held out of the model training. Once aggregating articles to the country level, the relationship between the number of articles classified by human raters to use data in a specific country and the machine learning model are even higher with a correlation of around 0.99. Applying this model to more than 1 million academic articles from the s2orc database from 2000-2020, the model predictions on articles using data per country are highly correlated with other estimates from the literature (Das et al. (2013), Porteous (2020), and National Science Board, National Science Foundation (2019)).

Based on the 1 million articles, the number of articles using data for 216 countries are calculated. 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. GDP per capita and population are both strong predictors of academic output. Combined the two explain around 74% of the cross country variation in academic output.

A countries SPI overall score is a statistically significant predictor of academic output using data, even conditional on population and GDP per capita. A particular area related to data use by academia are the data sources available for a country. For example, the availability of a populaiton census in the past 10 years is associated with a 0.6% increase in academic output using data. The availability of 2 or more household surveys on income/consumption in the past 10 years for a country is associated with an around 0.5% increase in papers using data. Availability of agriculture surveys and geospatial data at the Admin 1 level are also predictive.

The remainder of the paper clarifies our approach for measuring data use by academia and discusses our data sources. Section four discusses the model performance of the natural language processing approach. Section 5 discusses the empirical results. Section 6 includes the conclusions.

# 2 Empirical 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.

Second, whether the paper makes use of data will be classified. A supervised machine learning approach is used, where the team creates a precise definition of data use, manually labels 3500 publications randomly selected, and then trains a model on this data. In order convert the raw text of the title and abstract of the paper into a format suitable for the natural language processing model, the BERT (bidirectional Encoder Representations for transformers) model (Devlin et al. (2018)) is used to encode raw text to a numeric format suitable for predictions. BERT is pre-trained on a large corpus comprising the Toronto Book Corpus and Wikipedia. After encoding the raw text, PyTorch (Paszke et al. (2019)) is then used to produce a model to classify the articles based on the labeled data from human raters using the Mechanical Turk service. After training the model using BERT and PyTorch, out of sample predictions of whether or not an article uses data are possible using the trained model using the entire corpus of 1 million articles from s2orc.

Following the training of the model, 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 2000. 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 per capita, and a third measure we will consider is the share of papers on a country that use data.

# 3 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) | Data Use (1=yes) | Country Identified (1=yes) | Articles | Share of Articles |
| --- | --- | --- | --- | --- | --- |
| Business | 0.44 | 0.64 | 0.31 | 28,571 | 2.8 |
| Economics | 0.21 | 0.68 | 0.30 | 62,241 | 6.0 |
| Medicine | 0.04 | 0.85 | 0.10 | 840,920 | 81.0 |
| Political Science | 0.58 | 0.33 | 0.37 | 26,185 | 2.5 |
| Psychology | 0.25 | 0.70 | 0.15 | 44,191 | 4.3 |
| Sociology | 0.10 | 0.33 | 0.29 | 35,640 | 3.4 |

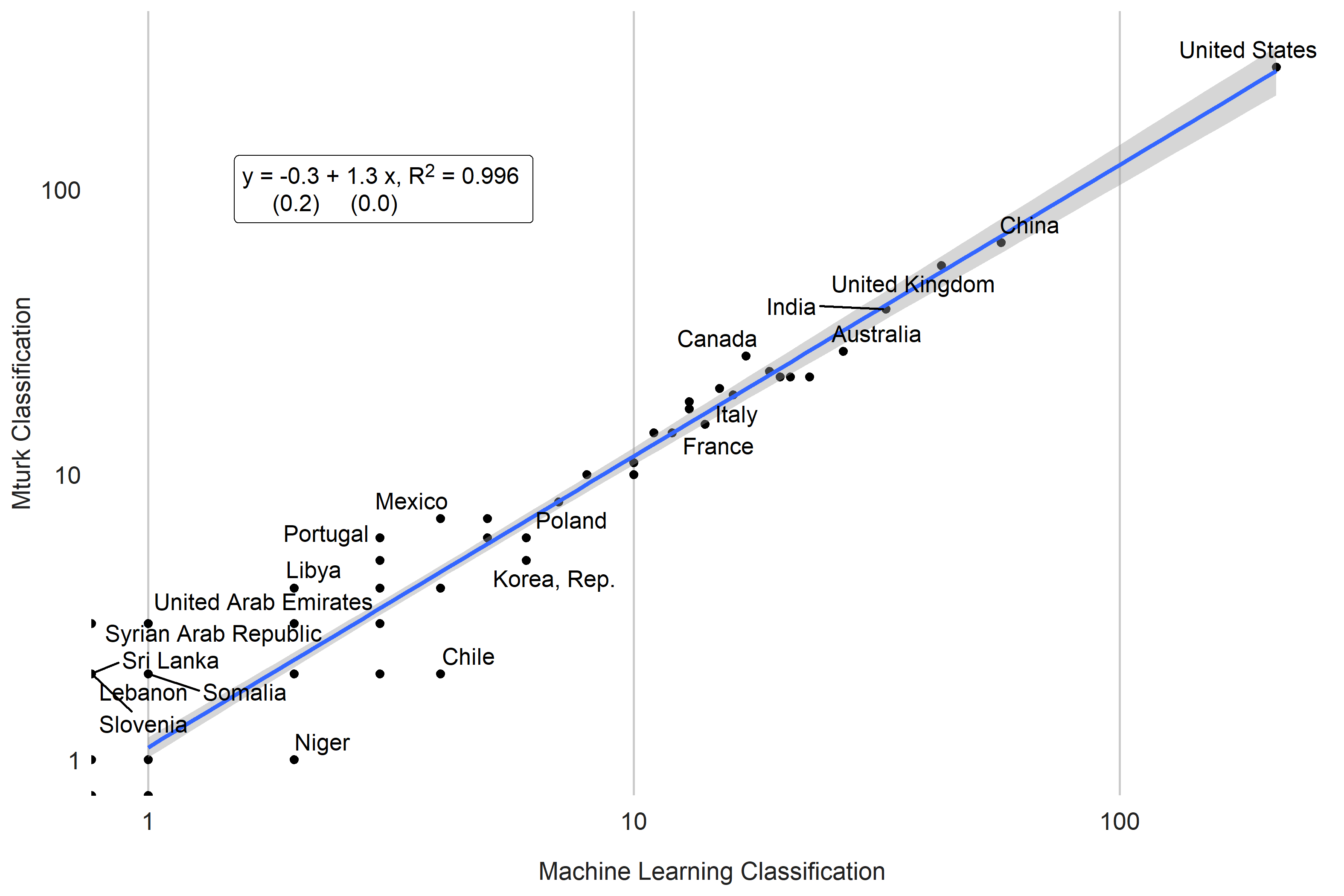
# 4 Model Performance

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.

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. From this set of 3,500 articles, 900 are then fed to the machine learning model, from which the country of study and data use status are identified.

After training a natural language processing model on around 900 hand coded articles using Amazon Mturk workers, the model was able to predict whether an article made use of data with 87% accuracy compared to a set of articles held out of the model training.

Among a holdout set of around 2,600 articles, which are not used in the model training, 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 in the figure below. The Pearson between the human raters and the NLP predictions is 0.998. The figure below shows the scatterplot between the human classifications and the NLP classifications.

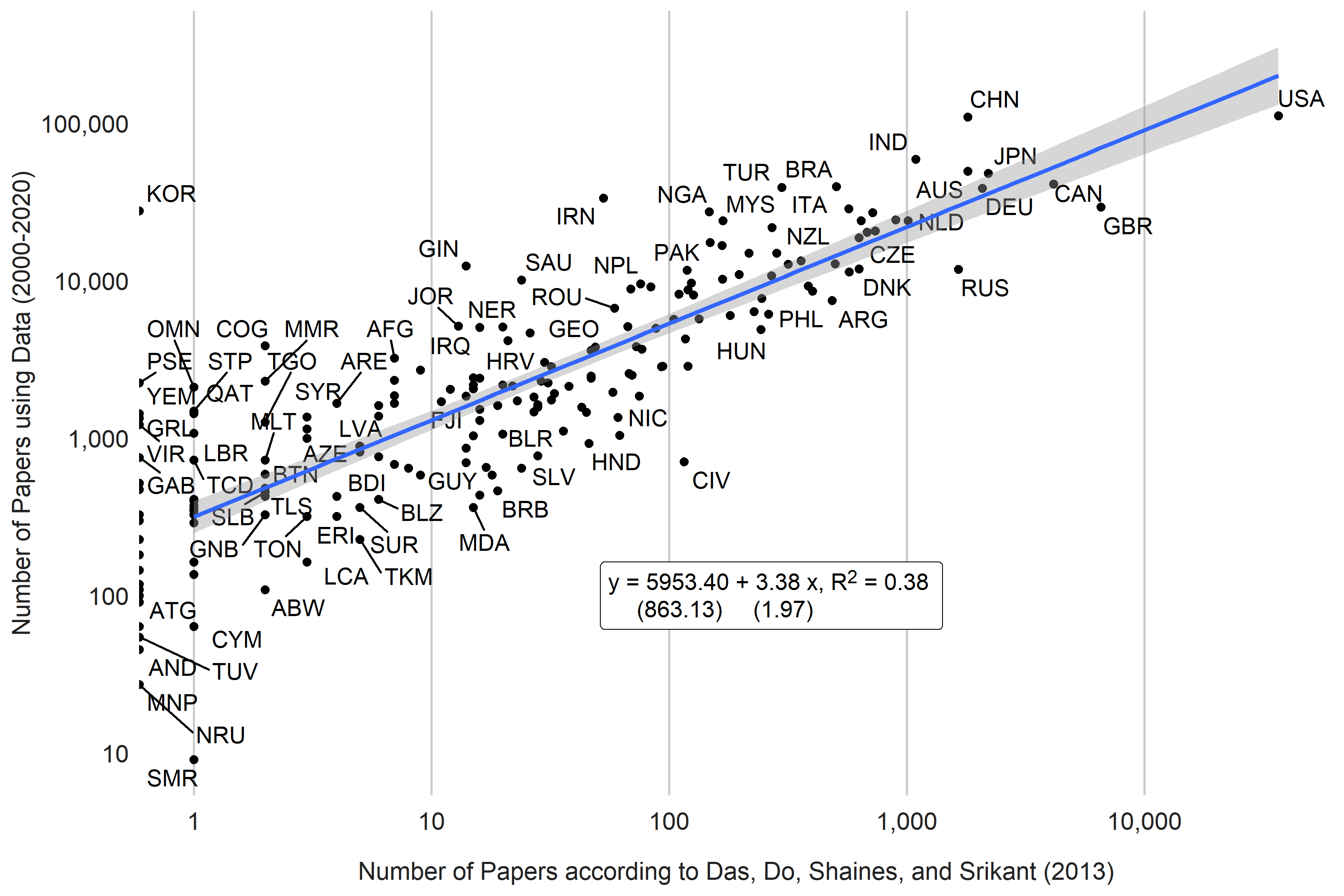
Figure. Comparison of Human Classifications of Data Use to NLP Predictions 

## 4.1 Comparison to Previous Literature

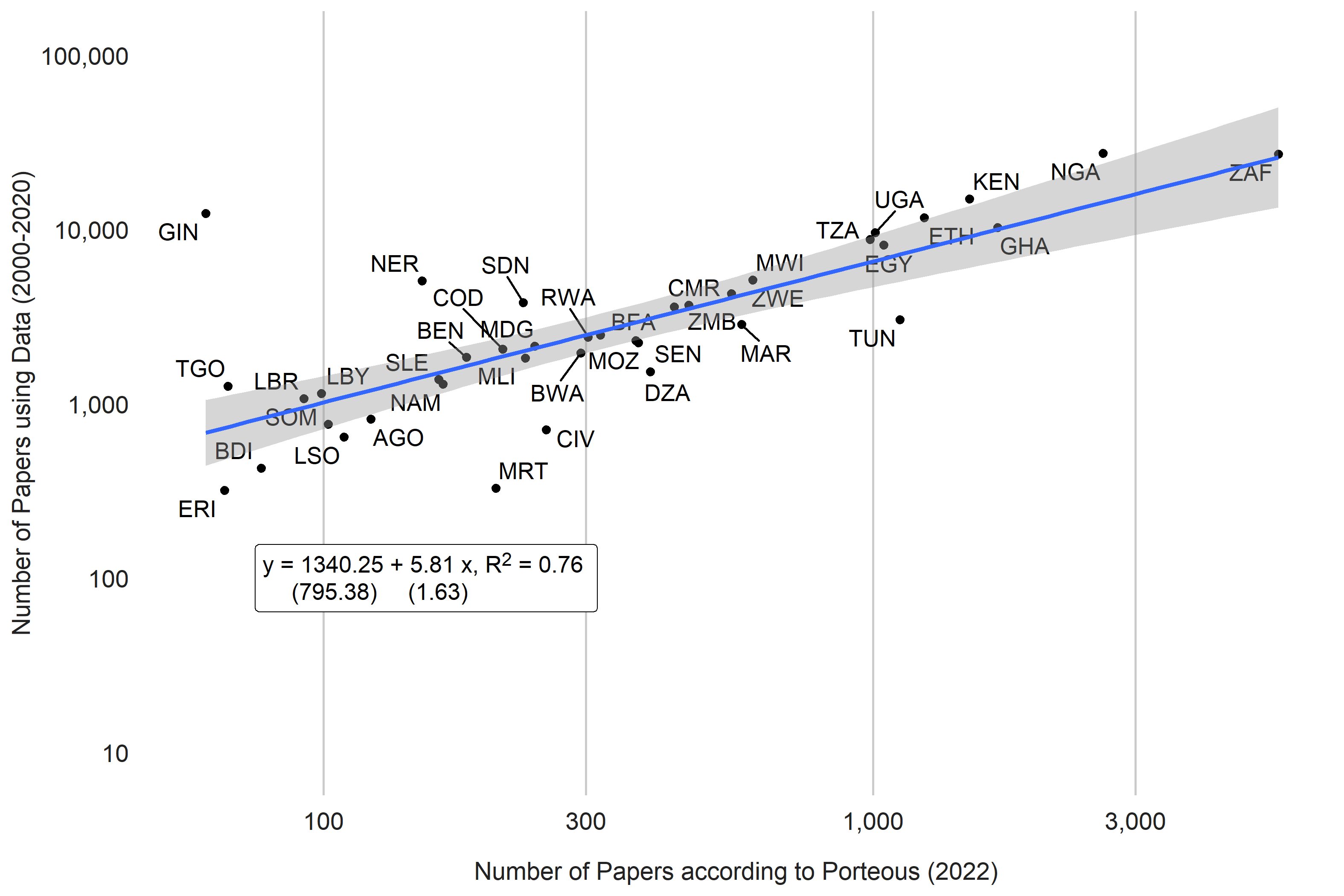
After applying the natural language processing model to around 1 million articles from the s2orc corpus, the estimated number of articles produced using data per country are then compared to previous estimates in the literature.

Das et al. (2013) use a corpus of more than 76 thousand empirical economics papers published between 1985 and 2005 to rank the academic output of countries using the EconLit database. While the estimates from our approach using the s2orc database do not exactly overlap, because of differences in the years covered and the subjects included (Das et al. (2013) consider only economics papers, while this study includes papers additionally from medicine, political science, sociology, and psychology), the correlation between country output is still 0.62. Estimates from Das et al. (2013) are taken from Table A3 of their paper.

Figure. Comparison to Number of Academic Articles in Das et al. (2013)

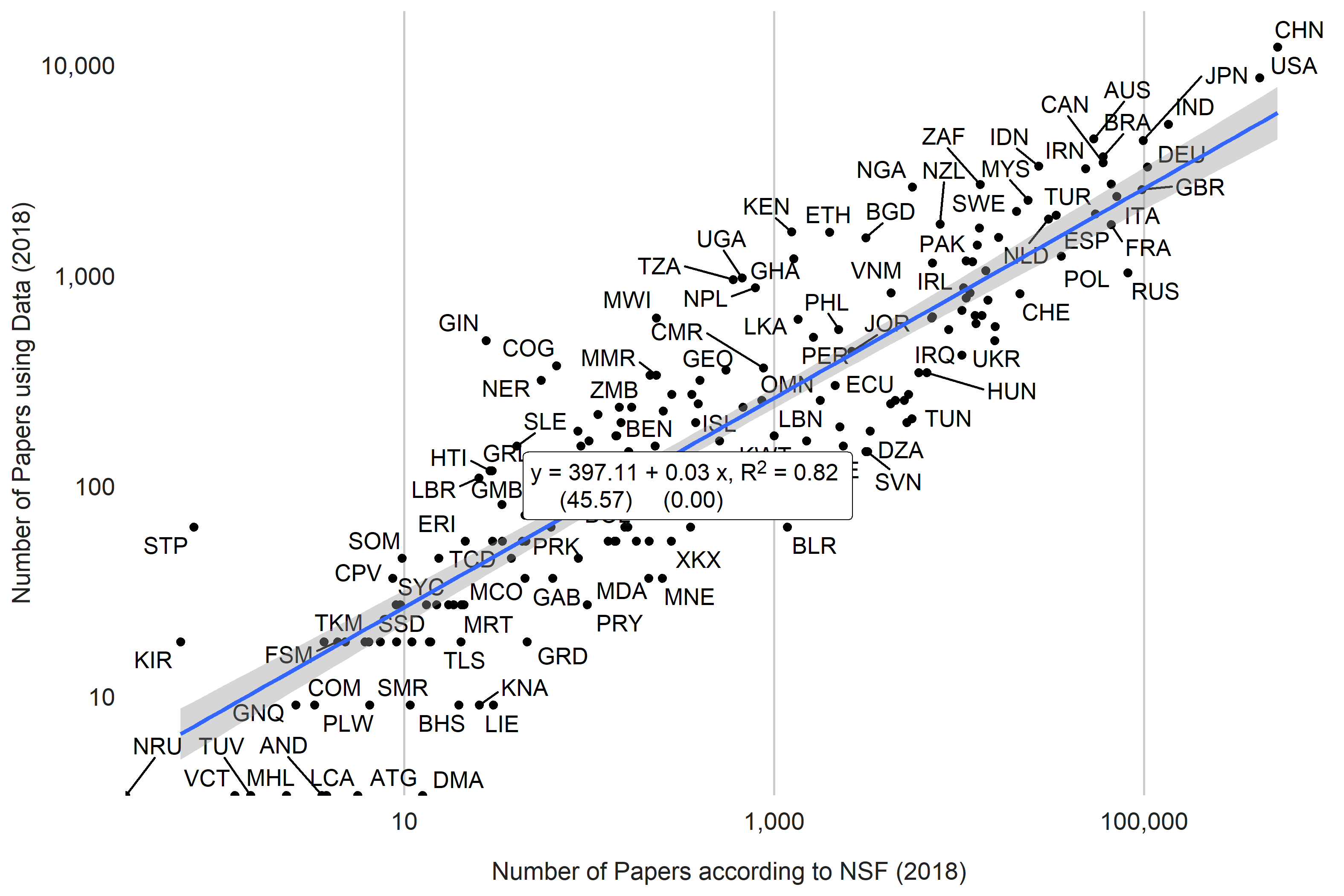


Additionally, Porteous (2020) examines the production of economics journal articles from 54 African countries between 2000 and 2019 using the EconLit database. The correlation between country rankings using the approach in this paper and that of Porteous (2020) is 0.87. Estimates from Porteous (2020) are taken from Table 3 in that paper and include all journal publications from 2000 to 2019 (column 2 of Table 3).

Figure. Comparison to Number of Academic Articles in Porteous (2020) 

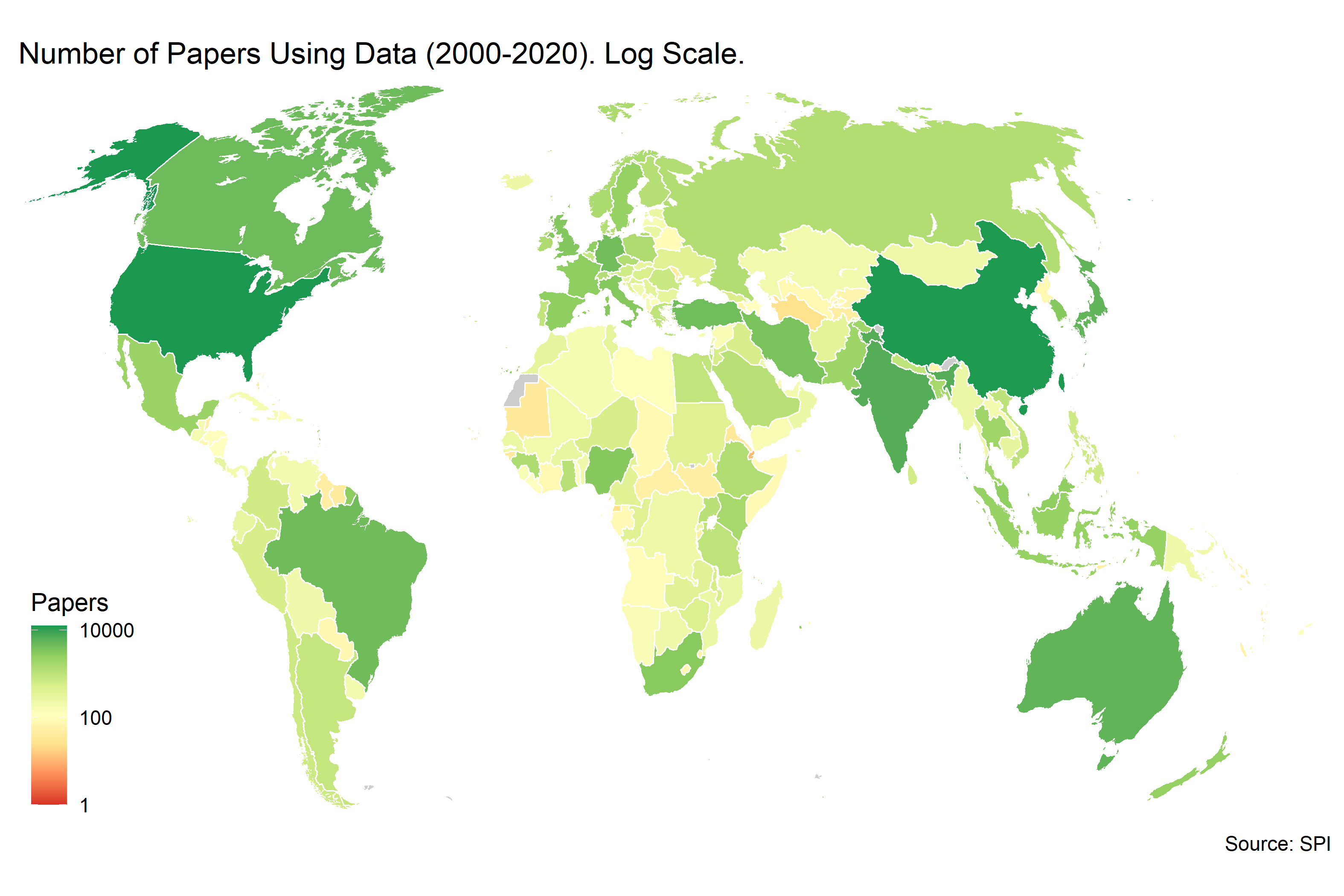
A third comparison is to the number of scientific and technical journal articles produced by the National Science Foundation (National Science Board, National Science Foundation (2019)). The number of scientific and engineering articles published in the following fields: physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and earth and space sciences. The NSF considers article counts from a set of journals covered by Science Citation Index (SCI) and Social Sciences Citation Index (SSCI). Data are accessed from the WDI (indicator IP.JRN.ARTC.SC).

Data from the NSF are papers produced in the year 2018. Likewise, estimates from this paper’s NLP model are for the year 2018 only. The correlation between the two estimates are around 0.9.

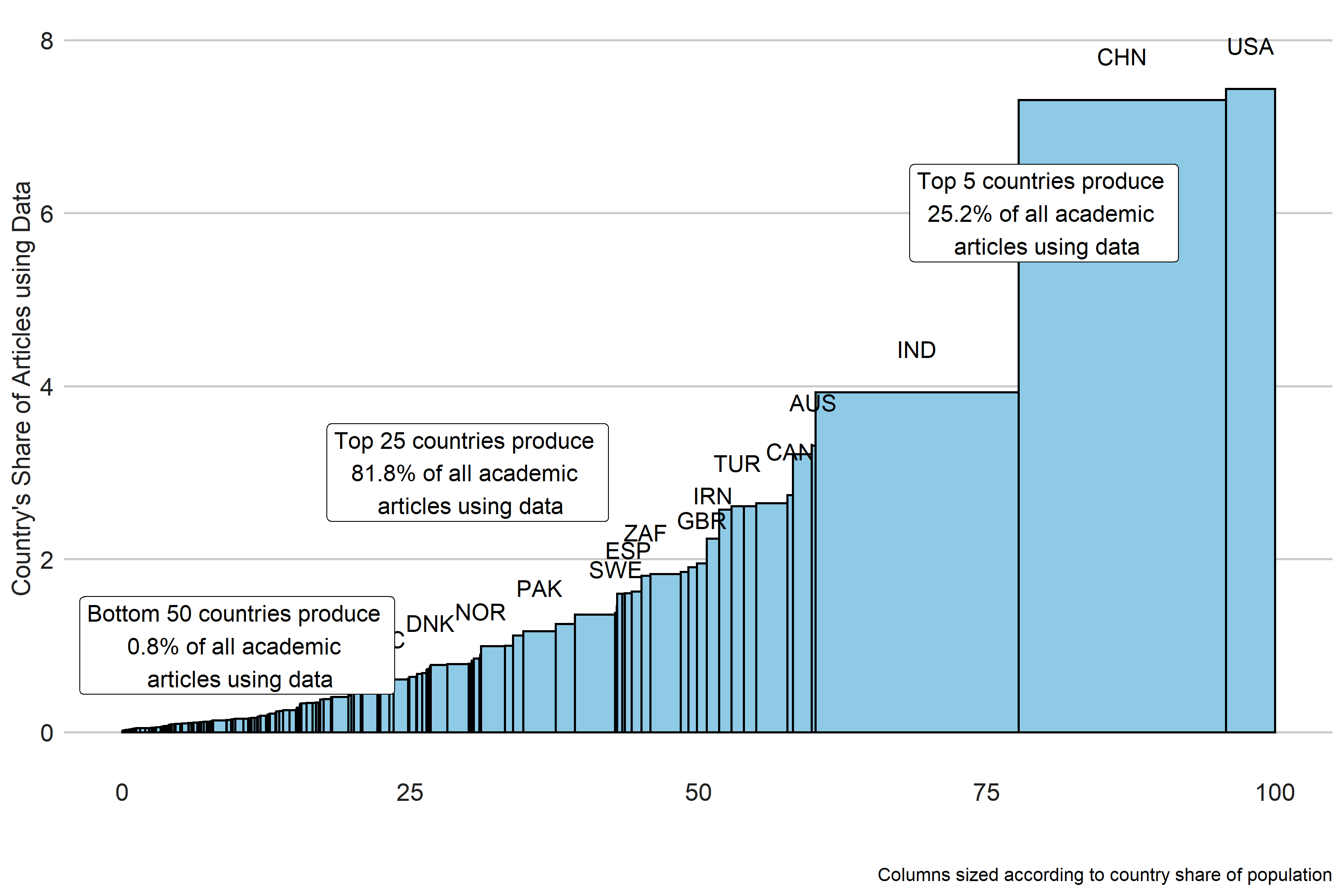
Figure. Comparison to Number of Academic Articles in NSF database of scientific and technical articles. 2018. 

# 5 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.

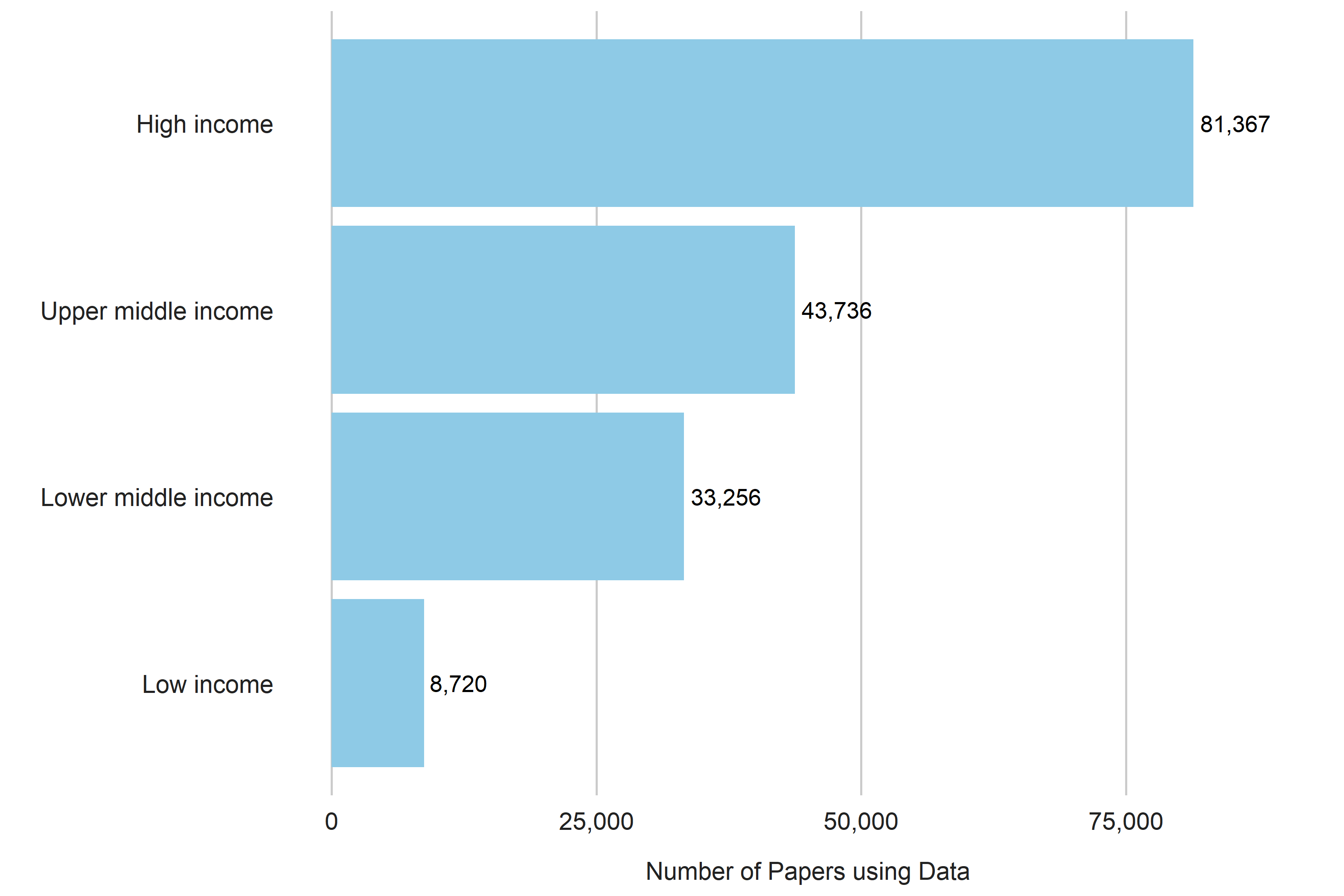
Figure. Number of Articles using Data by Country (2000-2020) 

The top 5 countries make up more than 25% of all academic output using data. However, these countries make up more than 40% of the World’s population. The top 25 countries make up more than 80% of output, while the bottom 50 countries produce less than 1% of output. These shares do match roughly to the population size of each of these groups, with the top 25 countries making up around 80% of the World population and the bottom 50 countries making up less than 1% of the population.

Figure. Distribution of Academic Output Across Countries. 

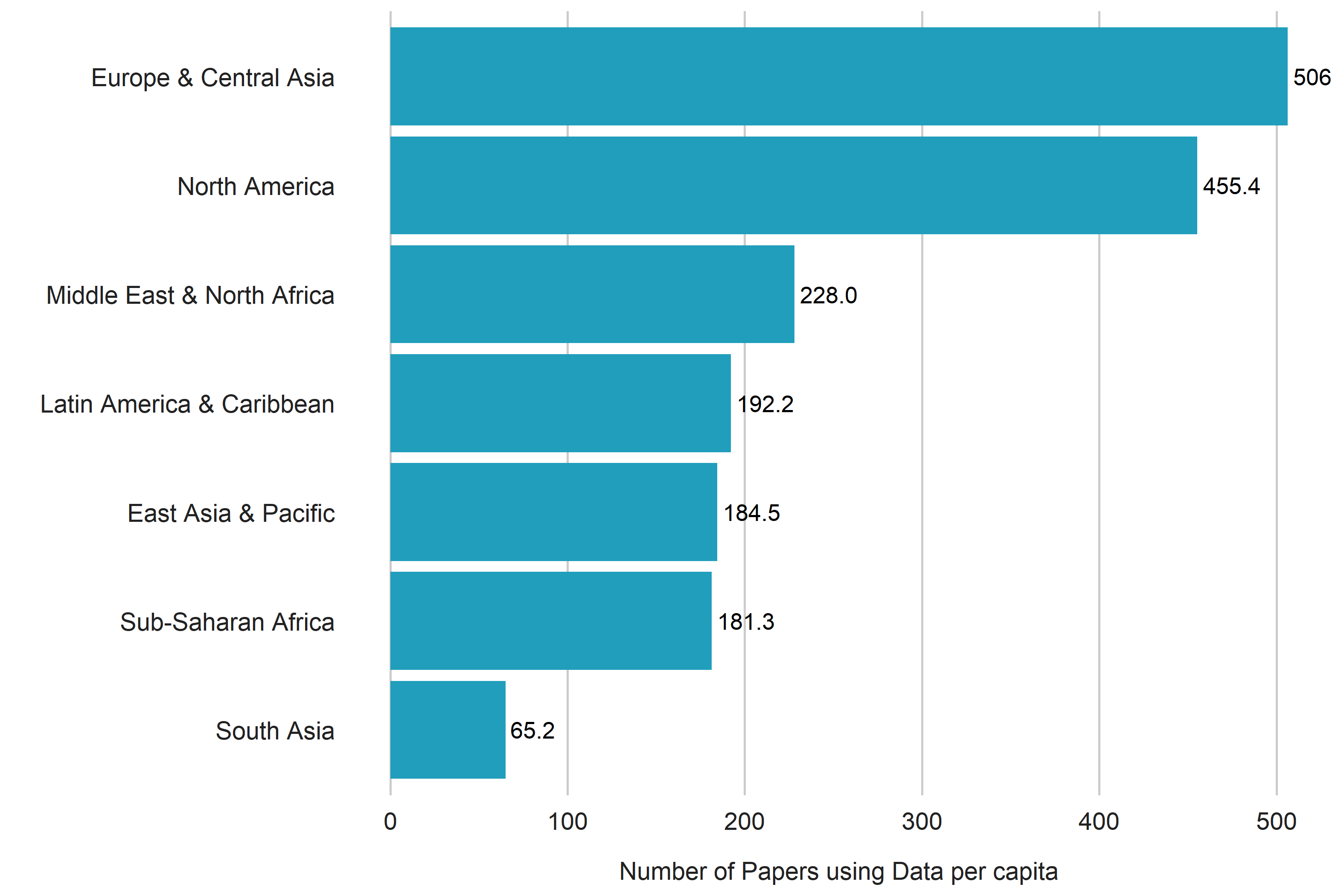
There is much more imbalance when looking at the income groups. Among income groups, high income countries produced nearly 50% of all papers using data from 2000-2020, despite only making up around 17% of the world’s population. Despite making up around 1/3rd of the world population, low income countries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

Figure. Total Number of articles using data by income

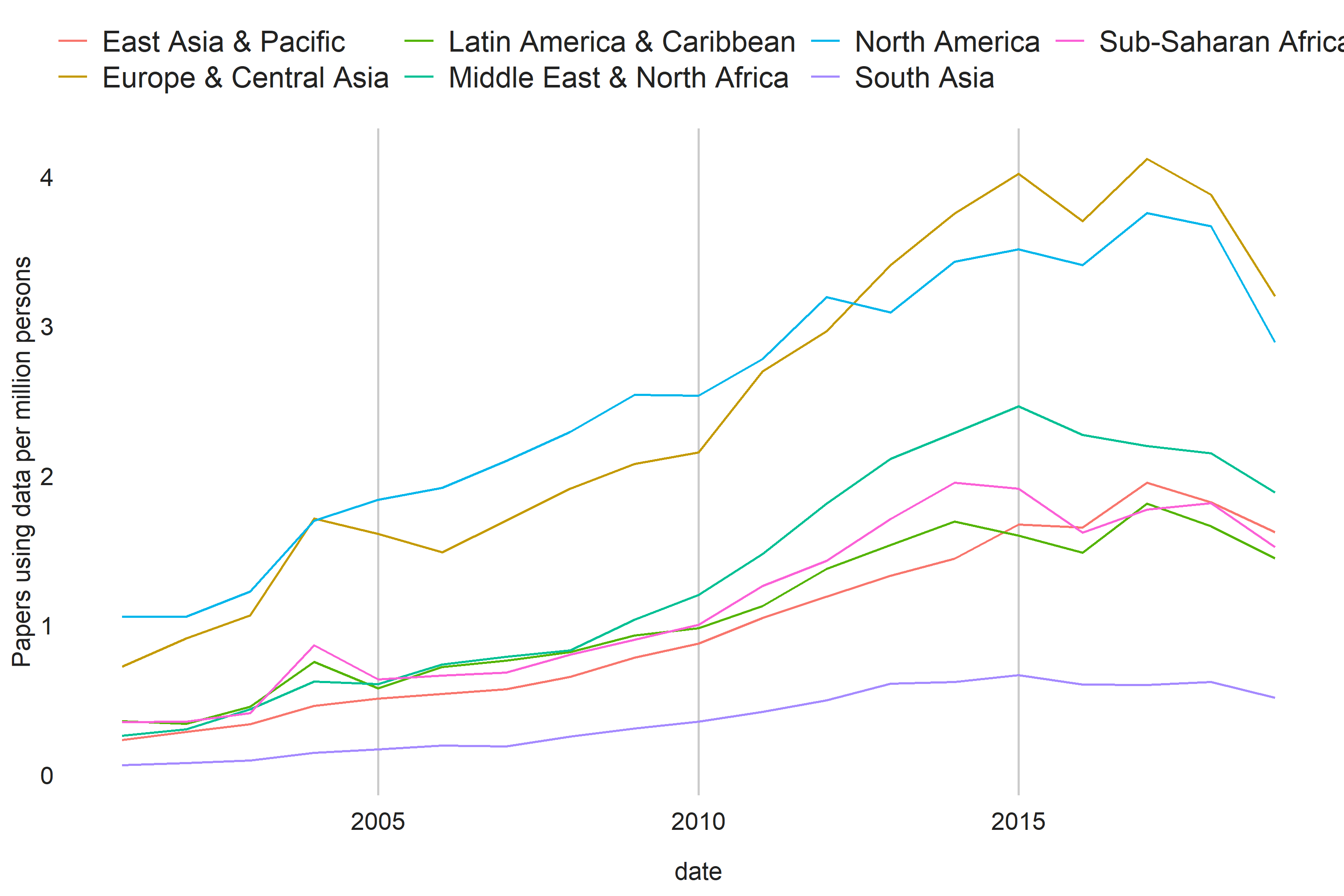


The numbers in the figure below shows the number of articles using data produced per million people in each region. When looking by region, Europe and Central Asia had the largest number of articles using data produced on a per capita basis with around 470 articles produced per million. North America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



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.

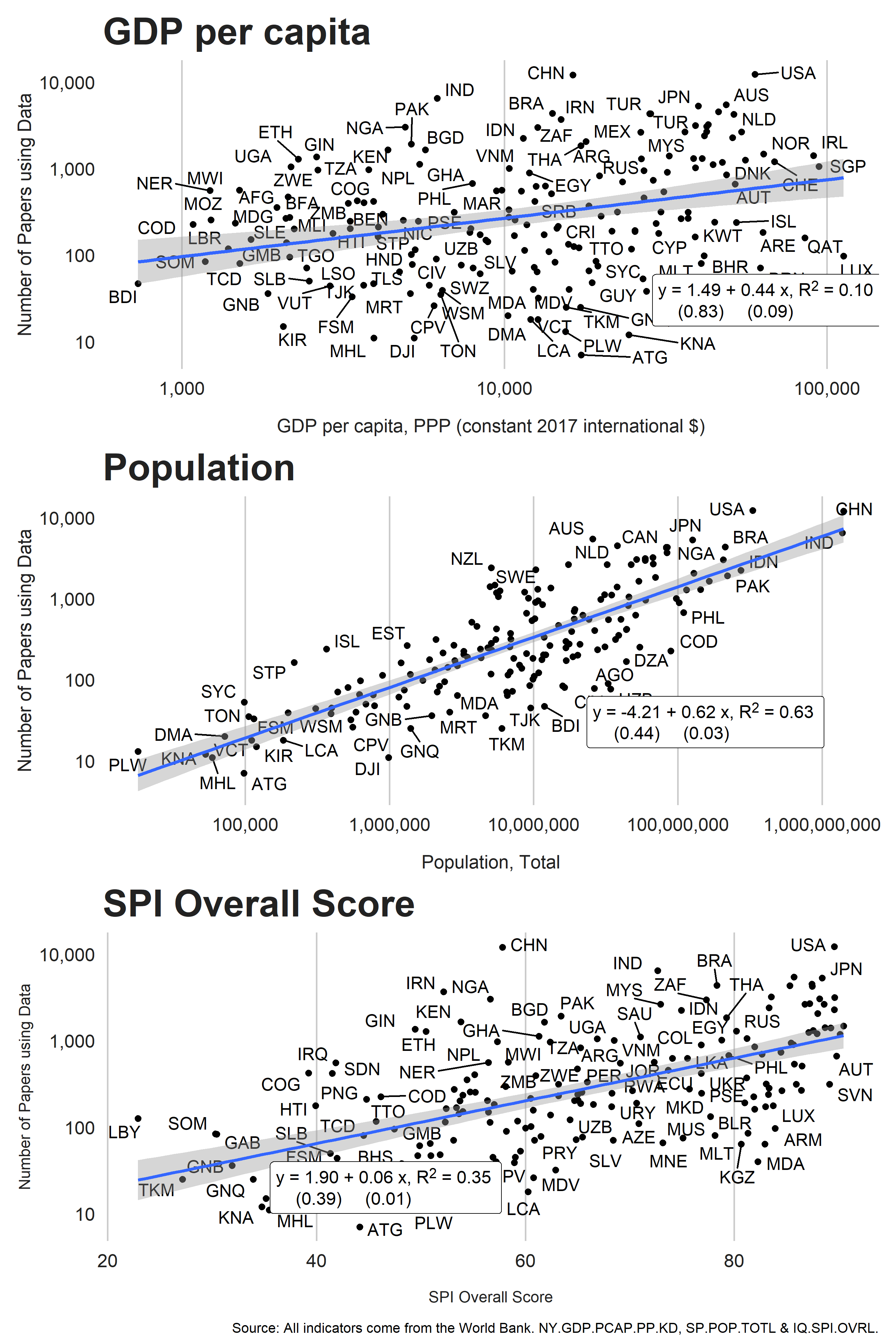
Figure. Articles Using Data per million persons over Time. 

There is a strong relationship between the number of articles produced per capita using data and GDP per capita, 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 per capita, the figures indicate an elasticity of 0.4, meaning that for instance a 10% increase in GDP translate into a 4% increase in academic articles using data. The elasticity with respect to population is slightly larger than of GDP per capita, with an elasticity of 0.6, and population alone explains 64% 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



The table below shows OLS regression coefficients from a regression of the log number of academic papers using data on log GDP per capita, log population, and the SPI scores. Log GDP per capita, population, and the SPI overall scores are all statistically significant at conventional levels. Log GDP per capita and Log Population combined explain close to 75% of the variation between countries in academic output. This is based on the R-squared from the linear regression. When the SPI overall scores are added around 76% of the variation in explained. The regressions indicate that a 10% increase in GDP per capita is associated with a 4.4% increase in academic output using data. A 10% increase in population is associated with a 6.2% increase in academic output.

The performance of a country’s statistical system also is associated with greater academic output. Conditional on GDP per capita and population, a ten point increase in SPI overall scores (on a scale of 0-100) translates into a 0.2% increase in academic output. The SPI pillar most strongly linked to academic output is the pillar on data sources, which cover the availability of recent censuses, surveys, academic data, and geospatial data..

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

|  | (1) | (2) | (3) | (4) | (5) |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 1.49+ | -4.21\*\*\* | 1.90\*\*\* | -7.76\*\*\* | -7.22\*\*\* |
|  | (0.83) | (0.44) | (0.39) | (0.77) | (0.99) |
| Log GDP per capita | 0.44\*\*\* |  |  | 0.31\*\*\* | 0.26\*\* |
|  | (0.09) |  |  | (0.07) | (0.09) |
| Log Population |  | 0.62\*\*\* |  | 0.58\*\*\* | 0.59\*\*\* |
|  |  | (0.03) |  | (0.03) | (0.03) |
| SPI Overall Score |  |  | 0.06\*\*\* | 0.02\*\*\* |  |
|  |  |  | (0.006) | (0.005) |  |
| SPI Data Use Score |  |  |  |  | -0.002 |
|  |  |  |  |  | (0.005) |
| SPI Data Services Score |  |  |  |  | 0.006 |
|  |  |  |  |  | (0.005) |
| SPI Data Products Score |  |  |  |  | 0.0008 |
|  |  |  |  |  | (0.008) |
| SPI Data Sources Score |  |  |  |  | 0.01\* |
|  |  |  |  |  | (0.006) |
| SPI Data Infrastructure Score |  |  |  |  | 0.004 |
|  |  |  |  |  | (0.003) |
| Num.Obs. | 179 | 179 | 179 | 179 | 179 |
| R2 | 0.100 | 0.633 | 0.348 | 0.784 | 0.789 |
| R2 Adj. | 0.095 | 0.631 | 0.345 | 0.781 | 0.780 |
| RMSE | 1.50 | 0.96 | 1.28 | 0.74 | 0.73 |
| Std.Errors | Heteroskedasticity-robust | Heteroskedasticity-robust | Heteroskedasticity-robust | Heteroskedasticity-robust | Heteroskedasticity-robust |
| Data from the World Bank's World Development Indicators (WDI) and SPI. WDI series codes include NY.GDP.PCAP.PP.KD, SP.POP.TOTL, IQ.SPI.OVRL,IQ.SPI.PIL1,IQ.SPI.PIL2, IQ.SPI.PIL3, IQ.SPI.PIL4, IQ.SPI.PIL5. Papers include all papers using data years 2000-2020.  \*\*\*=0.001 level  \*\*=0.01 level  \*=0.05 level  +=0.1 level | | | | | |

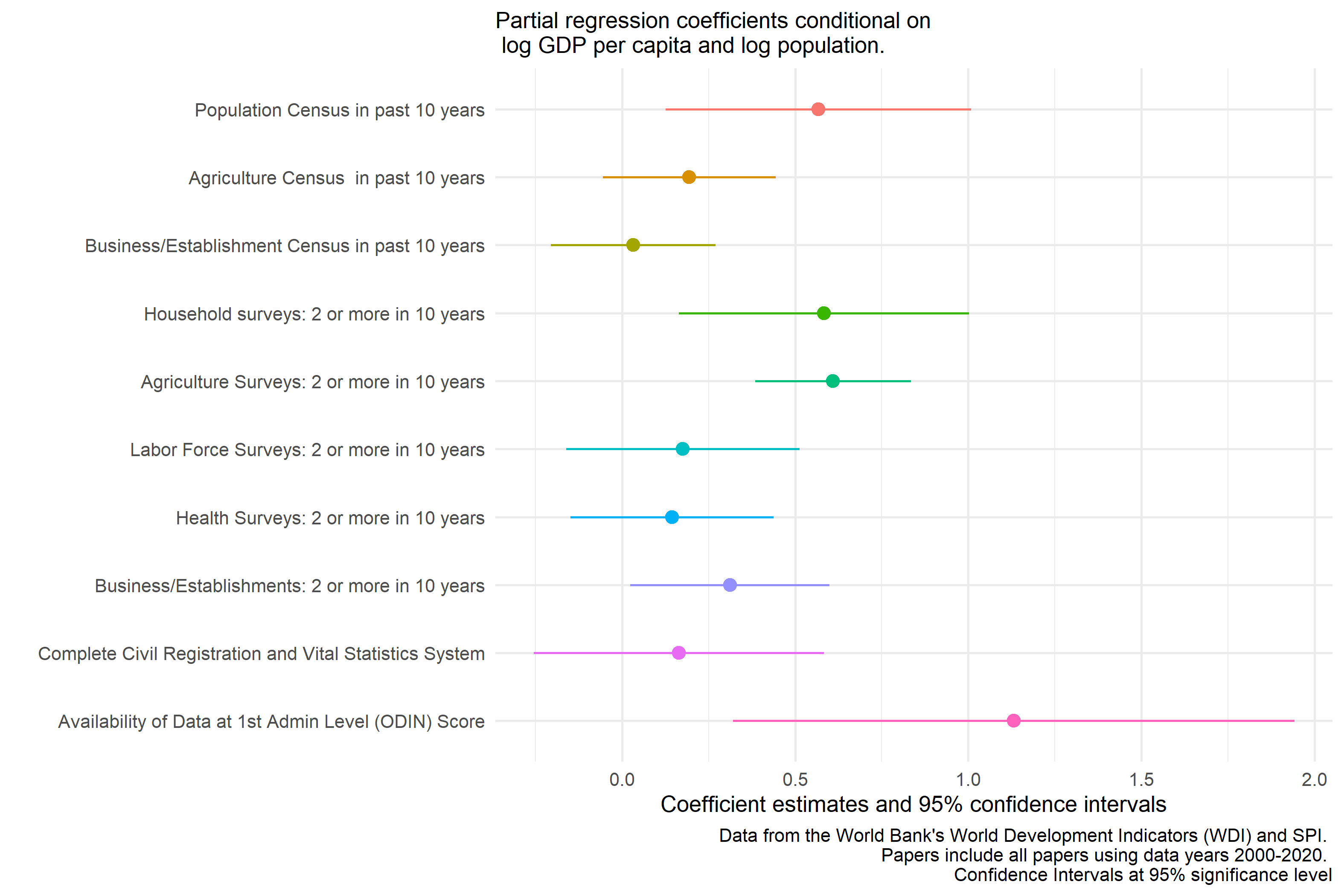
Table. Longitudinal Relationships between Number of Papers Using Data and Statistical Performance Scores

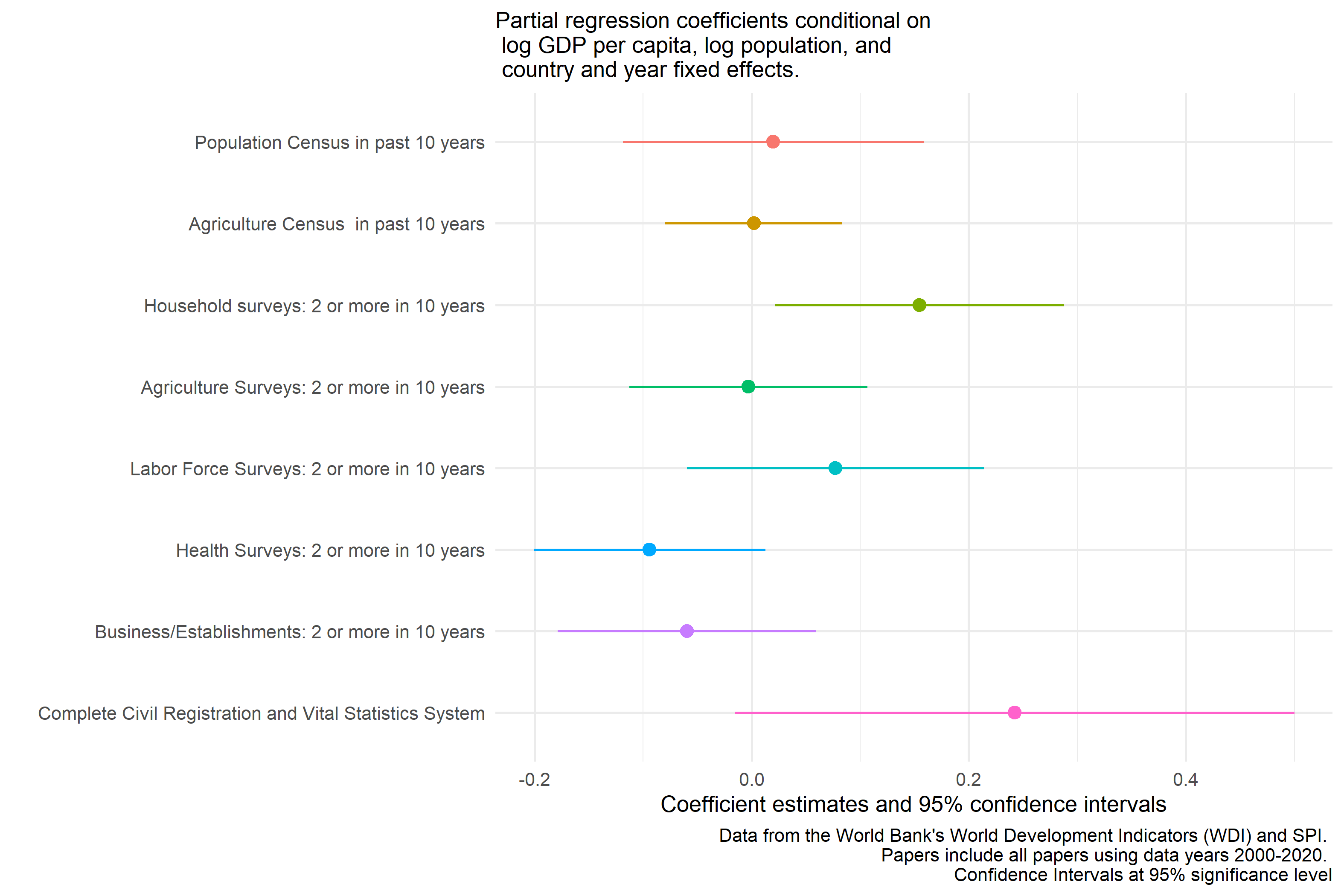
|  | (1) | (2) | (3) | (4) | (5) | (6) |
| --- | --- | --- | --- | --- | --- | --- |
| Log GDP per capita | 0.44\* |  |  | 0.58+ | 0.43\* | 0.38+ |
|  | (0.20) |  |  | (0.33) | (0.22) | (0.22) |
| Log Population |  | 0.19 |  | 1.51\* | 0.34 | 0.29 |
|  |  | (0.38) |  | (0.76) | (0.40) | (0.38) |
| SPI Overall Score |  |  | -0.003 | -0.002 |  |  |
|  |  |  | (0.004) | (0.004) |  |  |
| SPI Overall Score (Extended Series) |  |  |  |  | 0.005+ |  |
|  |  |  |  |  | (0.003) |  |
| SPI Data Use Score (Extended Series) |  |  |  |  |  | 0.002 |
|  |  |  |  |  |  | (0.001) |
| SPI Data Services Score (Extended Series) |  |  |  |  |  | 0.002 |
|  |  |  |  |  |  | (0.002) |
| SPI Data Products Score (Extended Series) |  |  |  |  |  | 0.004 |
|  |  |  |  |  |  | (0.003) |
| SPI Data Sources Score (Extended Series) |  |  |  |  |  | 0.003+ |
|  |  |  |  |  |  | (0.002) |
| SPI Data Infrastructure Score (Extended Series) |  |  |  |  |  | -0.0009 |
|  |  |  |  |  |  | (0.0007) |
| Num.Obs. | 2009 | 2009 | 682 | 682 | 2009 | 2009 |
| RMSE | 0.33 | 0.33 | 0.15 | 0.15 | 0.33 | 0.32 |
| Std.Errors | by: country | by: country | by: country | by: country | by: country | by: country |
| FE: country | X | X | X | X | X | X |
| FE: year | X | X | X | X | X | X |
| Data from the World Bank's World Development Indicators (WDI) and SPI. WDI series codes include NY.GDP.PCAP.PP.KD, SP.POP.TOTL, IQ.SPI.OVRL,IQ.SPI.PIL1,IQ.SPI.PIL2, IQ.SPI.PIL3, IQ.SPI.PIL4, IQ.SPI.PIL5. Papers include all papers using data years 2004-2019. SPI Extended Series data supplements SPI data with data from Statistical Capacity Indicator (SCI) to extend series back to 2004.  \*\*\*=0.001 level  \*\*=0.01 level  \*=0.05 level  +=0.1 level | | | | | | |

Using more detailed data from the Statistical Performance Indicators it is possible to assess how the availability of specific data sources relates to academic output. The figure below shows the partial linear regression coefficient for the availability of each of the 10 data sources considered by the SPI, conditional on log GDP per capita and log population. The ten data sources include: population census, agriculture census, business/establishment census, household consumption/income survey, agriculture survey, labor force survey, health survey, business/establishment survey, civil registration and vital statistics system (CRVS), and geospatial data at the Admin 1 (usually province/state) level.

The regression estimates indicate that the availability of a population census in the past 10 years is associated with an almost .6% increase in papers using data This estimate is significant at the 5% level. The availability of 2 or more household surveys measuring income/consumption in the past 10 years is associated with around a 0.5% increase in papers using data. This is also sifnificant at the 5% level. The availability of agricultural surveys is also a significant predictor at the 5% level. The share of indicators available at the Admin 1 level, calculated by Open Data Watch for their Open Data Inventory (ODIN) is associated with a 1% increase in academic output using data.

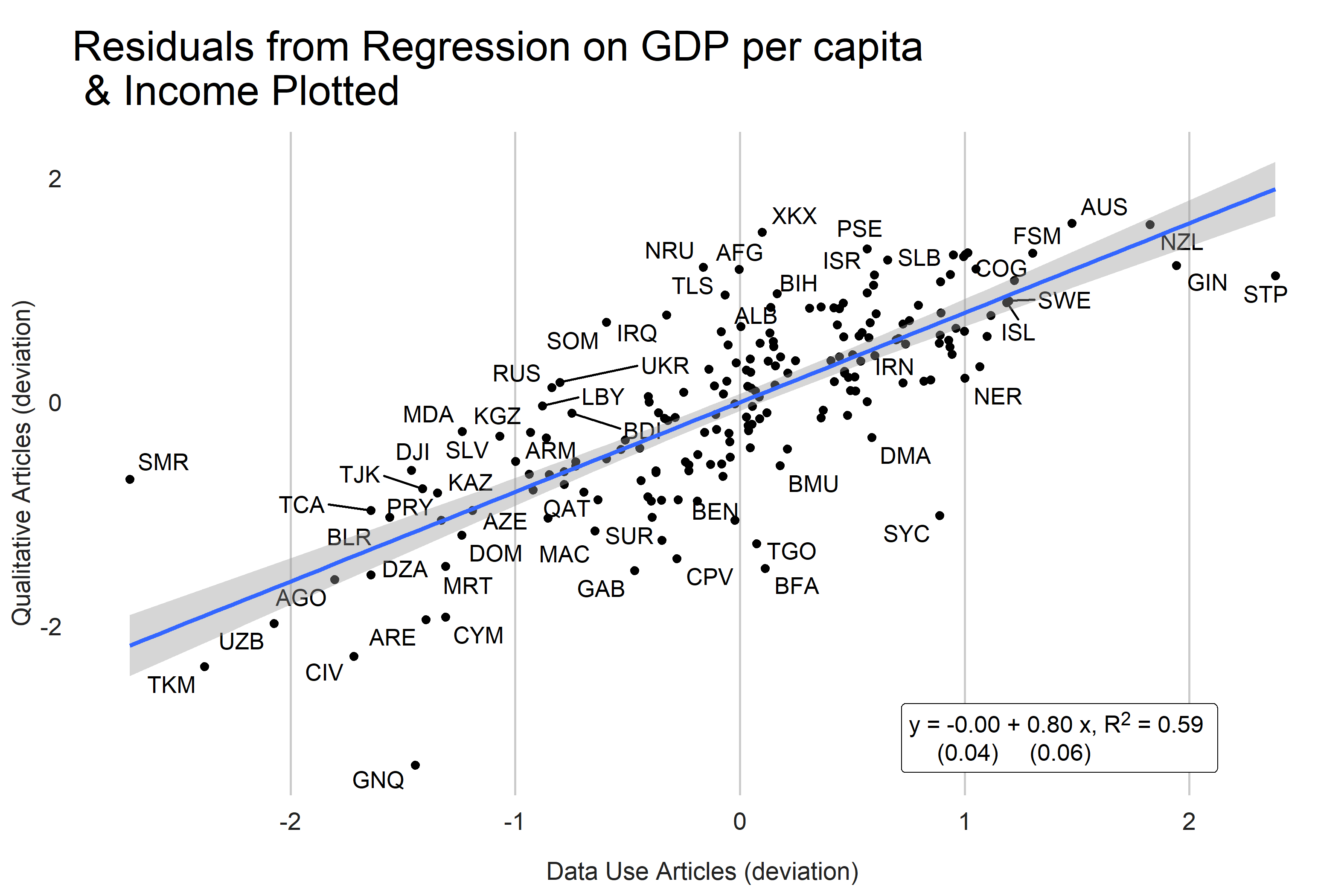
Figure. Relationships between Number of Papers Using Data and Availability of Data Sources.



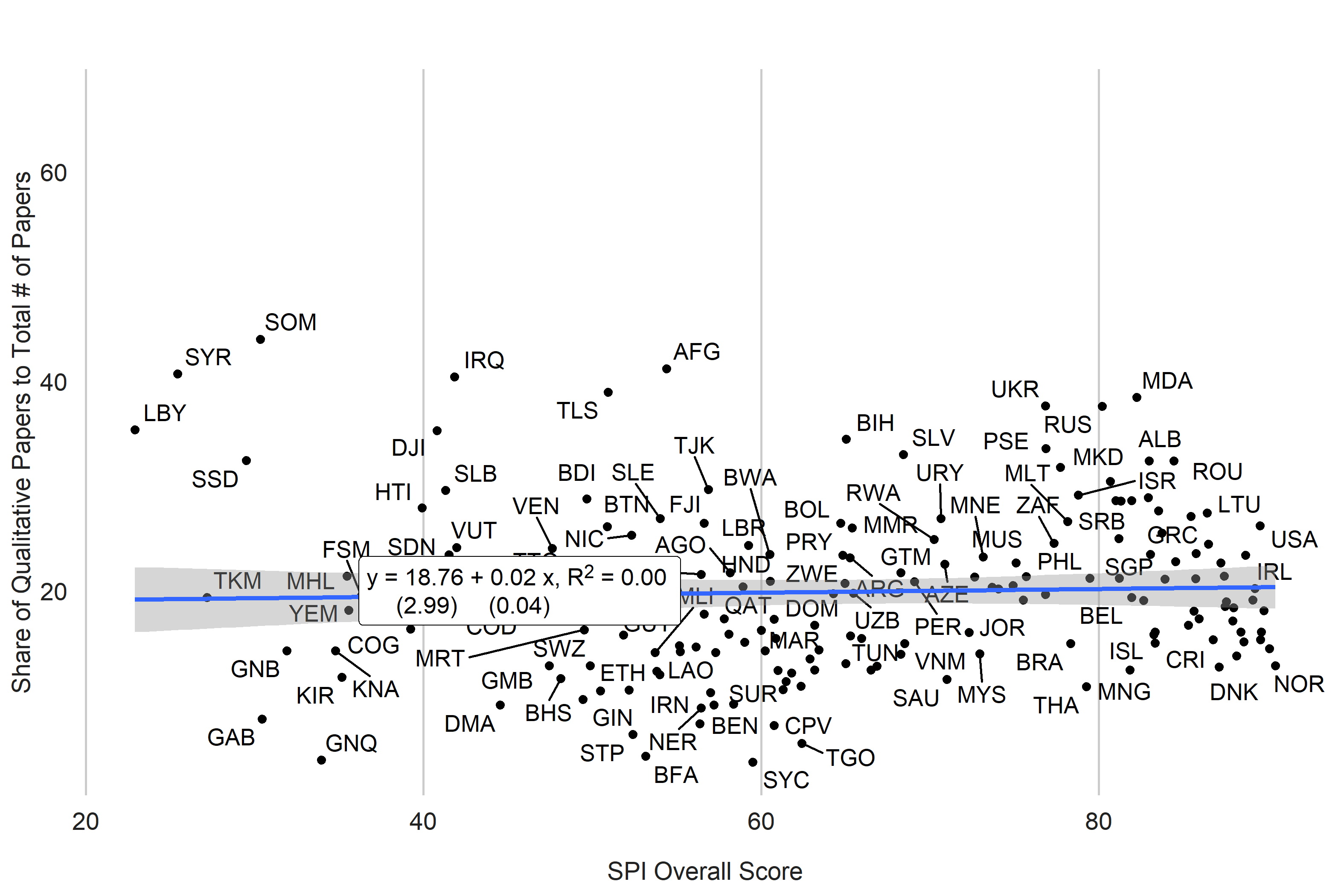


## 5.1 Qualitative versus Quantitative Papers

One may wonder whether countries with low levels of quantitative academic output (using data) compensate by producing more qualitative research. There does not appear to be evidence that this is the case. If this were the case, it might be expected that countries that under-produce quantitative articles then over-produce qualitative articles. The figure below shows the opposite to be the case. Countries that tend to over-produce quantitative articles also tend to over-produce qualitative articles, conditional on GDP per capita and population. The figure shows the residuals from the regression of the log number of quantitative, then qualitative, articles on log GDP per capita and population.

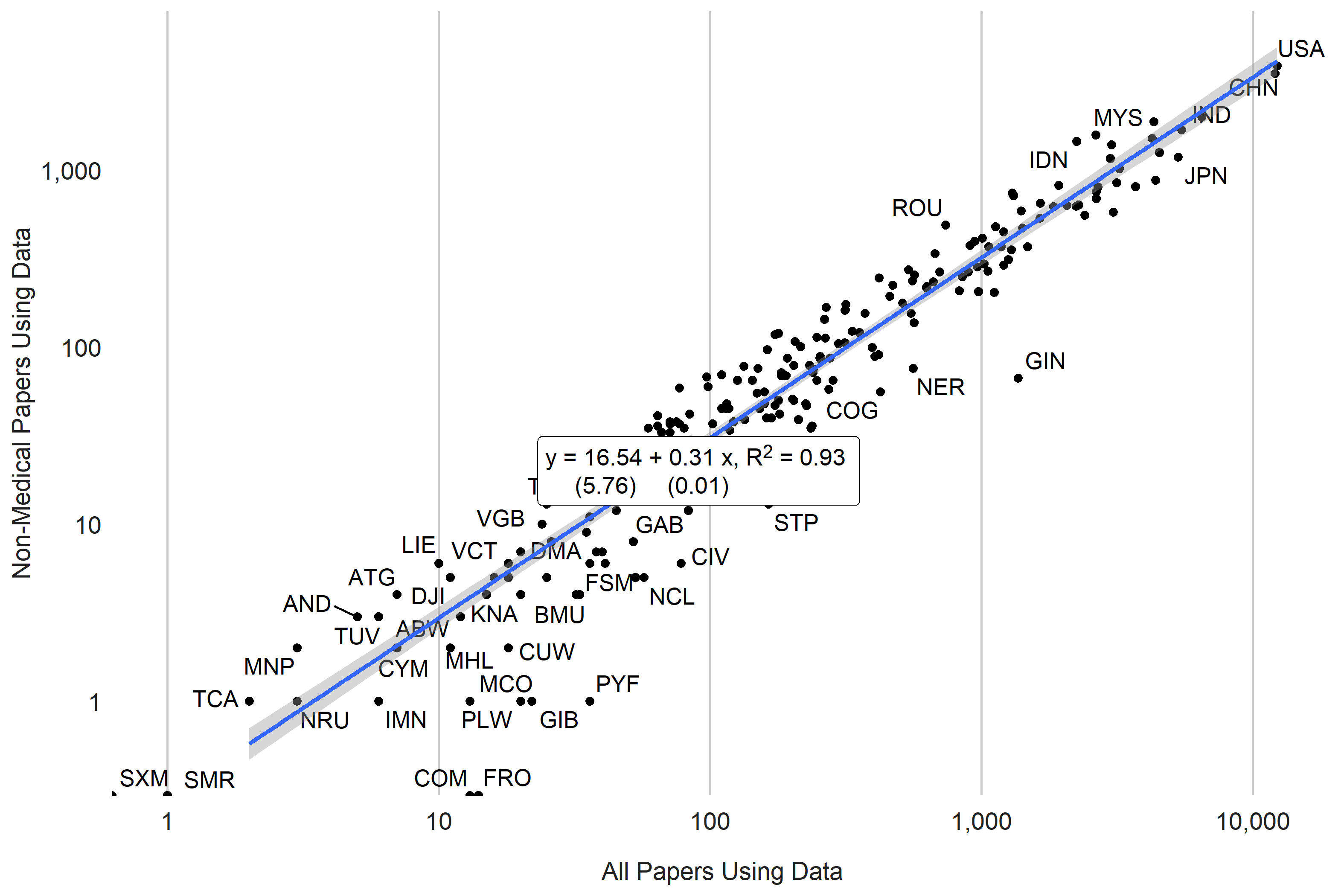
Figure. Comparison of the Quantity of Qualitative vs Quantitative Papers. 2000-2020. 

Another test is whether countries with weaker statistical systems tend to produce a higher share of qualitative research. Again, this would be expected if there is compensation, where places with weaker data systems produce more qualitative work to compensate. This also does not seem to be the case. There is almost no relationship between SPI overall scores and the share of qualitative papers to total papers produced overall.

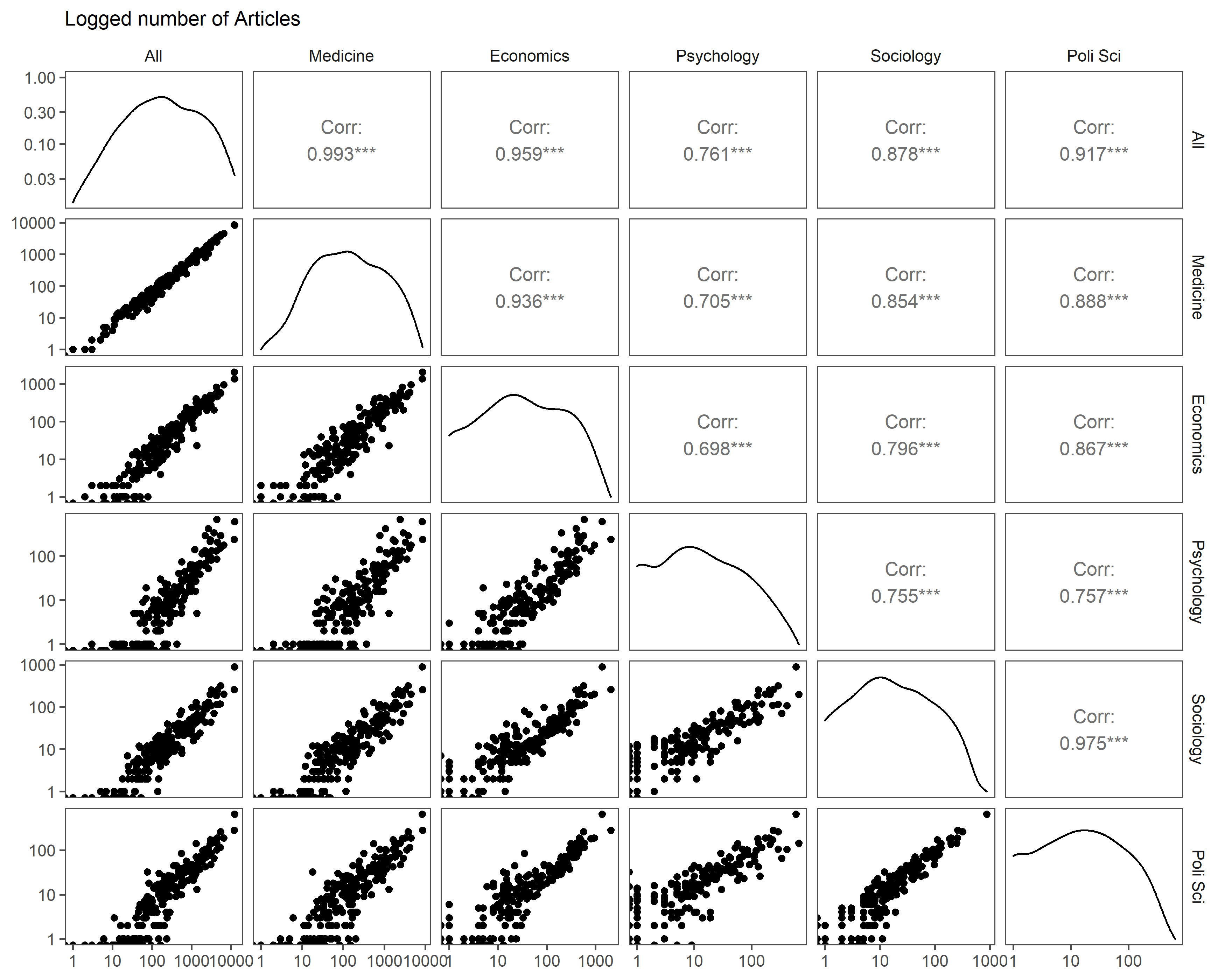
Figure. Comparison of Share of Qualitative vs SPI Overall Scores. 2000-2020. 

## 5.2 Sensitivity Checks

As noted previously, medical articles make up around 81% of all articles in our corpus. A concern could be that medical articles then receive too much weight in our final calculations of country scores. As a check, an alternative measure of the production of articles using data, which drops all medical articles, will be compared to the totals of the number of articles usind data described above.

Figure. Comparison Between All Papers in Corpus Using Data and Non-Medical Papers Using Data 

There are differences in academic output between subjects. The chart below shows the relationships between the number of articles per country using data for all subjects and for each subject respectively between 2000 and 2020. In all cases, academic output using data is strongly correlated across subjects. The correlation between the number of medical articles using data and economics is around 0.94. The subjects with the greatest correlation in articles using data is Political Science and Sociology with a correlation close to 0.98. Economics and Psychology have the lowest correlation (~0.75).

Figure. Correlation in Papers Using Data Across Subjects. 2000-2020. 

## 5.3 Annual 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: population size and GDP per capita. 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.

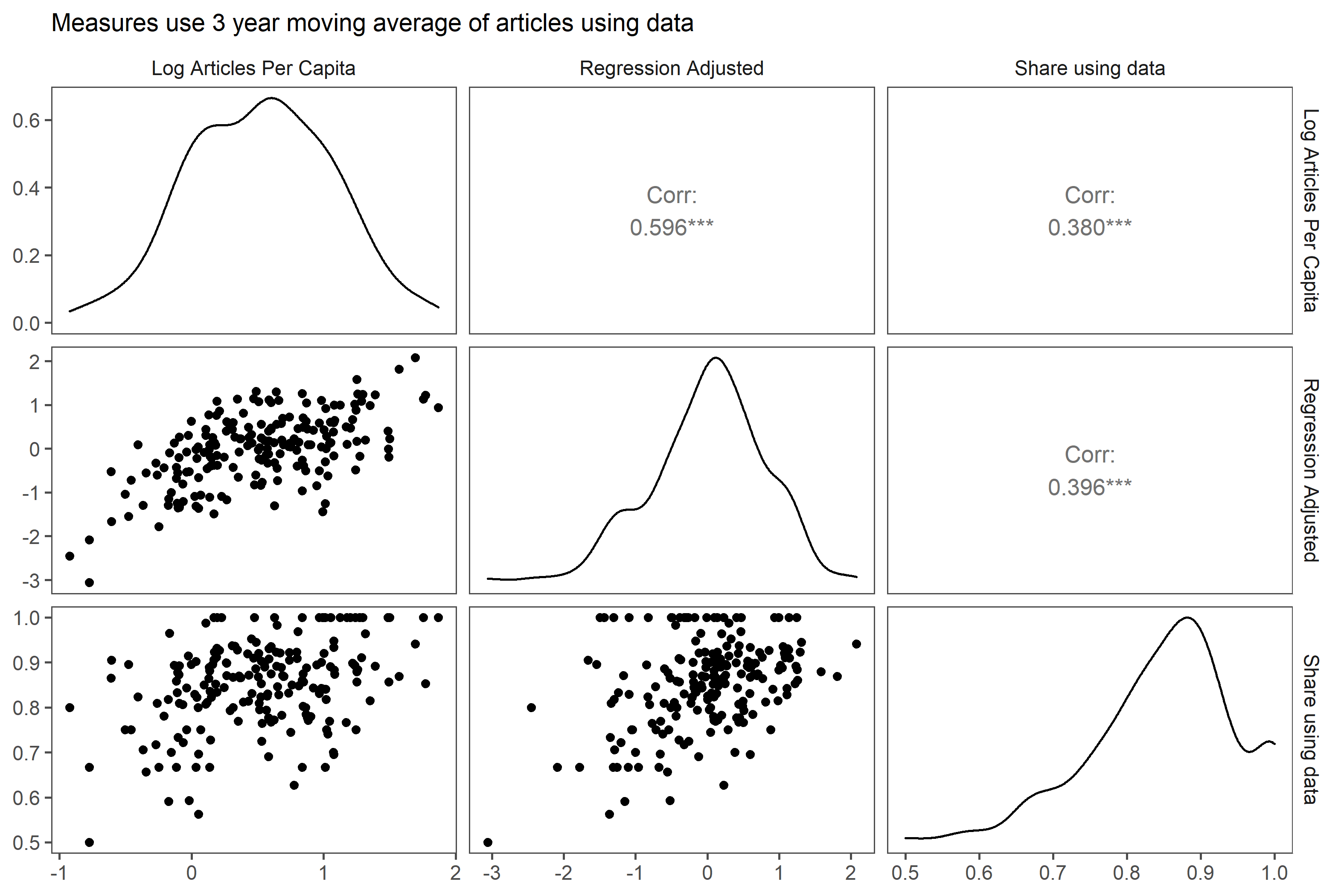
where

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.

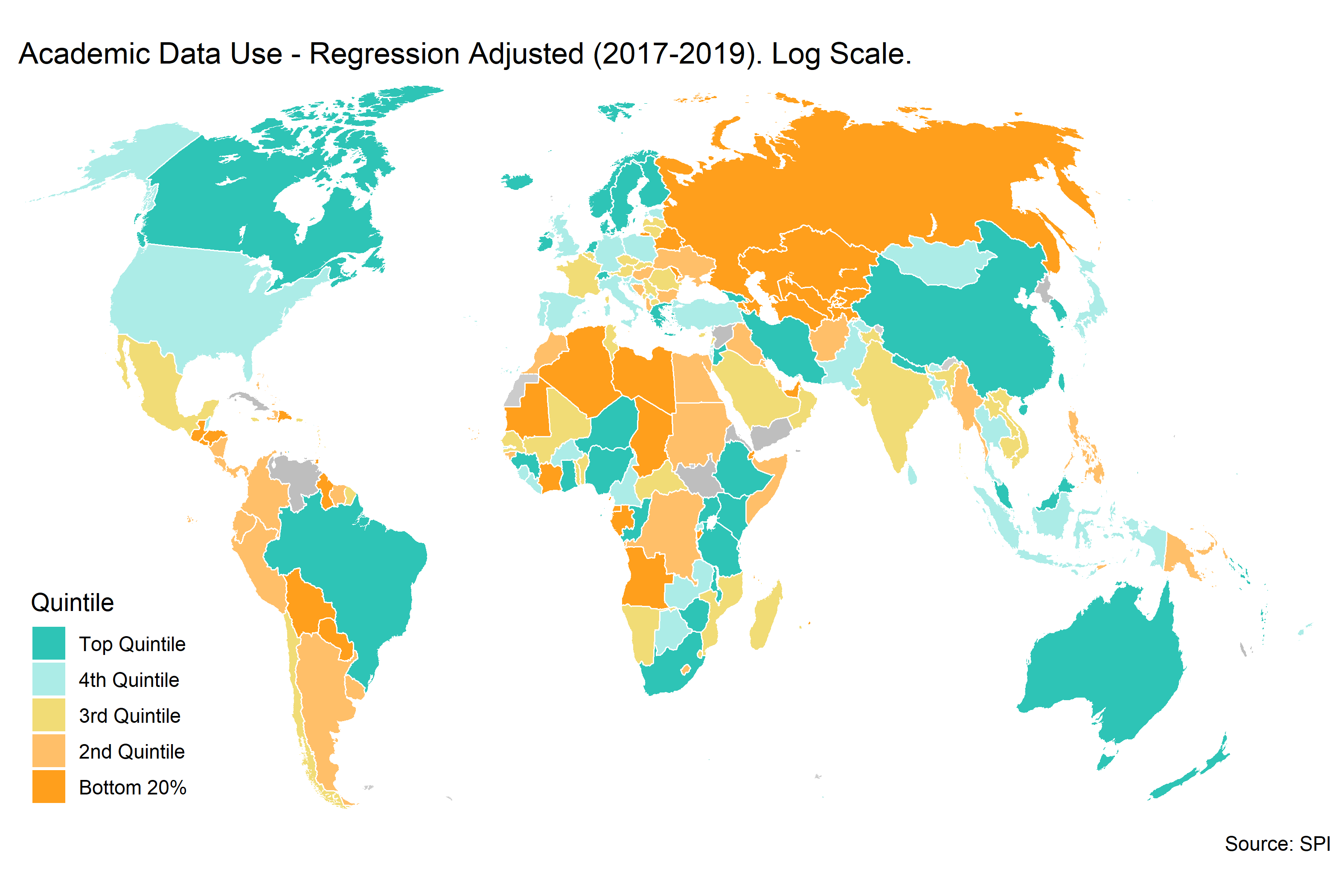
Because of volatility in the number of publications for a country in a particular year, especially for countries with relatively few publications, a 3 year moving average will be taken of the number of publications using data for each country.

The figure below shows the correlations between the three measures of data use described above: a per capita measure, a regression adjusted measure, and the share of articles using data. In the case of the per capita measure, the correlations shown are for the log number of articles per capita. All three measures are correlated with one another positively. The per capita measure and the regression adjusted measure have a pearson correlation of around 0.6, while the correlation with the share of papers using data for both the per capita and regression adjusted measure is close to 0.4.

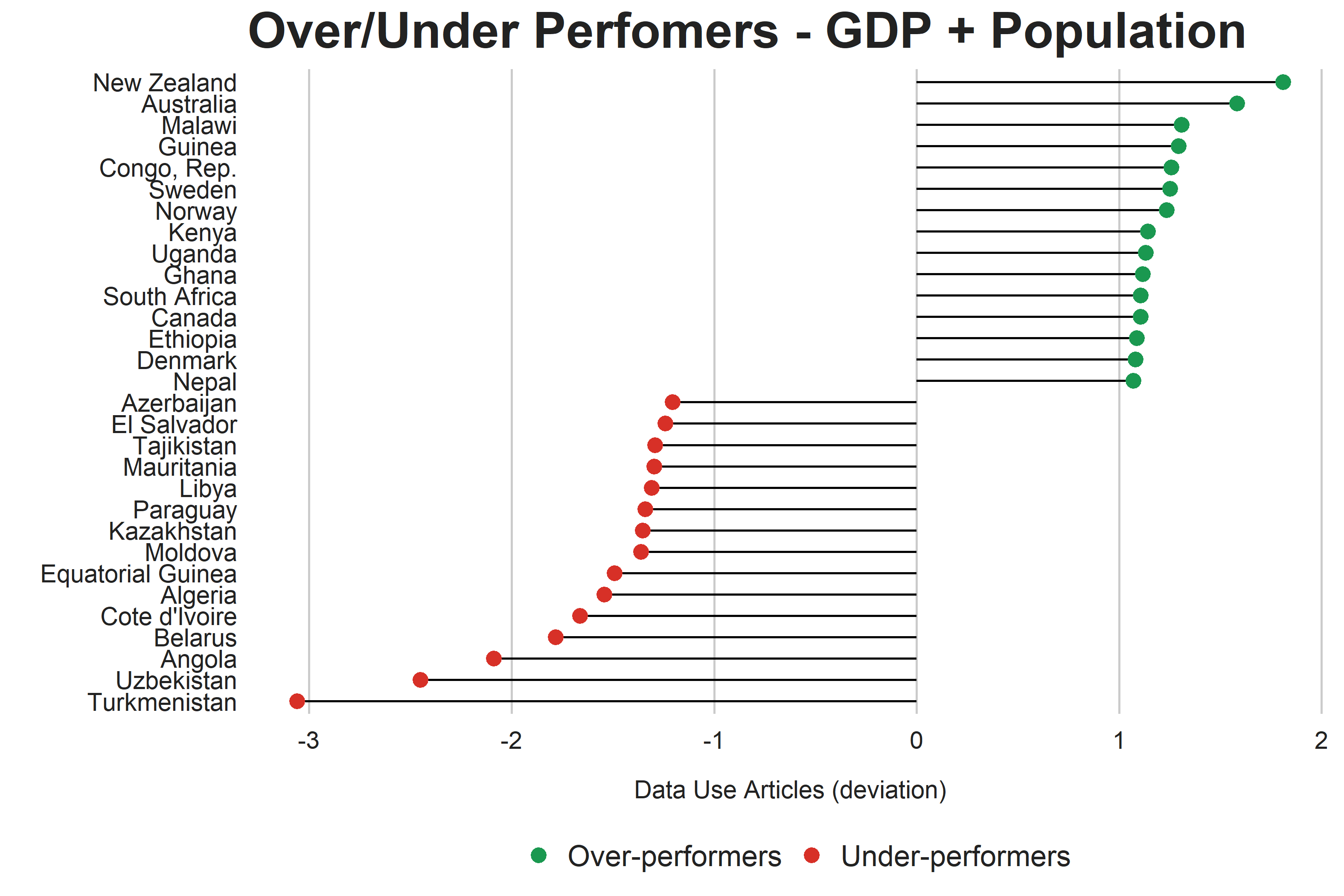
Figure. Country Scores under alternative scoring approaches 

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 .

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

Below we show the countries that over or underperform in terms of their GDP per capita and Population in 2020 compared to the number of papers using data. Only countries with more than 1 million persons will be included. This is a way to check for “outliers” in terms of academic output compared to income and population. This analysis is similar to that done in Das et al. (2013).



# 6 Conclusions

This paper suggests a methodology for measuring data use by academia by classifying academic articles using natural language processing tools. The NLP model can classify whether articles do or do not use data with 87% accuracy and match human rated classifications of academic output using data per country with a correlation of around 0.99. These measures also correlate strongly with alternative measures in the literature.

Applying the NLP model to a corpus of around 1 million academic articles a few key findings are revealed. High income countries produce a disproportionate share of academic output using data compared to their population. High income countries produce around 50% of all articles, while making up only 17% of the world population. Low income countries produce only 5% of academic output using data, while making up 1/3rd of the population. Second, GDP per capita and population are strong predictors of academic output, explaining around 75% of the variation between countries. Finally, statistical performance is a strong predictor of academic output using data, particularly in the case of data sources. Finally, a set of annual measures of academic output for 216 countries are produced.

With hundreds of millions spent on statistical capacity measures, the approach in this paper offers one metric for judging the effectiveness of these investments.

Next steps:

* Include final 10 million or so articles
* Further improve model performance
* Release as set of indicators for SPI

# 7 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.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv. <https://doi.org/10.48550/ARXIV.1810.04805>.

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>.

National Science Board, National Science Foundation. 2019. “Publications Output: US Trends and International Comparisons. Science & Engineering Indicators 2018.” *National Science Foundation*.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” arXiv. <https://doi.org/10.48550/ARXIV.1912.01703>.

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>.

# 8 Appendix

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

| Country | Academic Data Use per capita | Academic Data Use Regression Adjusted | Academic Data Use as Fraction of All Papers | Total Academic Publications using Data |
| --- | --- | --- | --- | --- |
| Sao Tome and Principe | 49.60 | 2.08 | 0.94 | 11 |
| New Zealand | 37.29 | 1.81 | 0.87 | 140 |
| Australia | 17.79 | 1.58 | 0.88 | 360 |
| Malawi | 3.08 | 1.31 | 0.95 | 50 |
| Guinea | 4.38 | 1.29 | 0.92 | 55 |
| Congo, Rep. | 6.88 | 1.26 | 0.86 | 32 |
| Sweden | 18.13 | 1.25 | 0.88 | 135 |
| Kiribati | 19.84 | 1.24 | 1.00 | 5 |
| Norway | 24.56 | 1.23 | 0.89 | 104 |
| Iceland | 59.17 | 1.22 | 0.85 | 18 |
| Kenya | 2.94 | 1.14 | 0.91 | 136 |
| Tuvalu | 57.20 | 1.14 | 1.00 | 2 |
| Uganda | 2.23 | 1.13 | 0.93 | 88 |
| Ghana | 3.85 | 1.12 | 0.93 | 111 |
| Canada | 9.59 | 1.11 | 0.84 | 310 |
| South Africa | 4.58 | 1.11 | 0.83 | 223 |
| Ethiopia | 1.55 | 1.09 | 0.93 | 187 |
| Denmark | 19.38 | 1.08 | 0.91 | 96 |
| Nepal | 3.22 | 1.07 | 0.92 | 69 |
| Iran, Islamic Rep. | 3.98 | 1.05 | 0.94 | 276 |
| Malaysia | 7.44 | 1.05 | 0.89 | 205 |
| Finland | 17.21 | 1.02 | 0.89 | 74 |
| Netherlands | 12.05 | 1.00 | 0.88 | 195 |
| Vanuatu | 13.34 | 1.00 | 1.00 | 7 |
| Ireland | 22.56 | 0.99 | 0.81 | 79 |
| Palau | 74.07 | 0.93 | 1.00 | 1 |
| Georgia | 10.39 | 0.91 | 0.84 | 36 |
| Micronesia, Fed. Sts. | 17.57 | 0.88 | 0.75 | 2 |
| Niger | 1.63 | 0.86 | 0.93 | 39 |
| Zimbabwe | 2.46 | 0.81 | 0.81 | 31 |
| Nigeria | 1.35 | 0.77 | 0.86 | 248 |
| Tanzania | 1.54 | 0.77 | 0.91 | 77 |
| Korea, Rep. | 5.51 | 0.73 | 0.92 | 248 |
| Solomon Islands | 6.97 | 0.70 | 0.88 | 3 |
| Jordan | 4.88 | 0.69 | 0.87 | 54 |
| Singapore | 16.54 | 0.67 | 0.90 | 89 |
| Switzerland | 12.09 | 0.65 | 0.87 | 91 |
| China | 0.99 | 0.63 | 0.90 | 1,308 |
| Greece | 7.37 | 0.63 | 0.78 | 72 |
| Brazil | 1.84 | 0.62 | 0.90 | 343 |
| Hong Kong SAR, China | 11.23 | 0.61 | 0.89 | 80 |
| Belize | 11.96 | 0.60 | 0.93 | 5 |
| Rwanda | 2.06 | 0.60 | 0.87 | 27 |
| Botswana | 9.26 | 0.59 | 0.83 | 18 |
| Fiji | 11.99 | 0.59 | 0.70 | 10 |
| Spain | 4.79 | 0.58 | 0.89 | 214 |
| Japan | 3.37 | 0.56 | 0.90 | 366 |
| Italy | 4.40 | 0.55 | 0.89 | 245 |
| Sierra Leone | 1.96 | 0.51 | 0.79 | 15 |
| Estonia | 14.82 | 0.50 | 0.77 | 17 |
| Portugal | 7.00 | 0.49 | 0.80 | 61 |
| United States | 2.70 | 0.49 | 0.81 | 748 |
| Germany | 4.01 | 0.47 | 0.87 | 286 |
| Tonga | 15.95 | 0.47 | 1.00 | 2 |
| Mongolia | 6.41 | 0.46 | 0.97 | 16 |
| Israel | 7.55 | 0.45 | 0.78 | 52 |
| Indonesia | 1.28 | 0.44 | 0.81 | 355 |
| Liberia | 2.03 | 0.44 | 0.94 | 9 |
| West Bank and Gaza | 3.98 | 0.44 | 0.77 | 19 |
| Bhutan | 10.05 | 0.43 | 0.92 | 5 |
| Croatia | 8.36 | 0.42 | 0.84 | 30 |
| United Kingdom | 3.83 | 0.41 | 0.78 | 222 |
| Seychelles | 30.73 | 0.41 | 1.00 | 2 |
| Zambia | 1.83 | 0.40 | 0.90 | 33 |
| Samoa | 11.84 | 0.38 | 0.70 | 3 |
| Gambia, The | 2.84 | 0.33 | 0.95 | 4 |
| Burkina Faso | 1.28 | 0.30 | 0.99 | 24 |
| Bangladesh | 0.94 | 0.30 | 0.91 | 121 |
| Slovenia | 10.53 | 0.29 | 0.75 | 19 |
| Pakistan | 0.81 | 0.27 | 0.89 | 180 |
| Thailand | 2.13 | 0.27 | 0.94 | 146 |
| Cameroon | 1.44 | 0.26 | 0.90 | 43 |
| Sri Lanka | 2.71 | 0.25 | 0.87 | 47 |
| Poland | 3.37 | 0.24 | 0.85 | 138 |
| St. Kitts and Nevis | 31.55 | 0.23 | 1.00 | 1 |
| Turkey | 2.36 | 0.23 | 0.87 | 169 |
| Kosovo | 5.96 | 0.23 | 0.63 | 9 |
| Turkiye | 2.33 | 0.22 | 0.87 | 169 |
| Brunei Darussalam | 20.77 | 0.20 | 0.96 | 8 |
| Namibia | 4.81 | 0.20 | 0.78 | 13 |
| Chile | 3.57 | 0.19 | 0.89 | 61 |
| Saudi Arabia | 3.84 | 0.18 | 0.91 | 134 |
| Grenada | 17.86 | 0.17 | 0.86 | 2 |
| Austria | 6.49 | 0.15 | 0.85 | 44 |
| Belgium | 5.74 | 0.15 | 0.85 | 65 |
| Marshall Islands | 11.34 | 0.14 | 1.00 | 1 |
| Serbia | 4.22 | 0.14 | 0.77 | 21 |
| Togo | 1.48 | 0.14 | 0.92 | 8 |
| France | 2.85 | 0.13 | 0.83 | 168 |
| Mozambique | 0.74 | 0.13 | 0.89 | 21 |
| Cyprus | 11.12 | 0.12 | 0.77 | 9 |
| Eswatini | 5.52 | 0.12 | 0.90 | 5 |
| Czech Republic | 4.99 | 0.10 | 0.87 | 45 |
| St. Vincent and the Grenadines | 15.07 | 0.10 | 1.00 | 2 |
| Czechia | 4.97 | 0.09 | 0.87 | 45 |
| India | 0.39 | 0.09 | 0.82 | 502 |
| Lithuania | 7.64 | 0.09 | 0.77 | 15 |
| Latvia | 8.01 | 0.09 | 0.78 | 13 |
| Senegal | 1.37 | 0.09 | 0.88 | 20 |
| Cambodia | 1.52 | 0.08 | 0.85 | 25 |
| Lao PDR | 2.65 | 0.07 | 0.92 | 20 |
| Slovak Republic | 5.44 | 0.06 | 0.83 | 19 |
| North Macedonia | 5.62 | 0.04 | 0.74 | 8 |
| Montenegro | 9.65 | 0.04 | 0.90 | 5 |
| Central African Republic | 1.12 | 0.04 | 0.80 | 3 |
| Romania | 3.27 | 0.02 | 0.79 | 49 |
| Barbados | 10.45 | 0.00 | 0.82 | 1 |
| Jamaica | 3.73 | 0.00 | 0.83 | 5 |
| Nauru | 30.97 | -0.01 | 1.00 | 0 |
| Benin | 1.36 | -0.02 | 0.89 | 15 |
| Vietnam | 1.08 | -0.02 | 0.90 | 104 |
| Lebanon | 3.21 | -0.03 | 0.87 | 21 |
| Puerto Rico | 6.26 | -0.03 | 0.92 | 19 |
| Mali | 0.92 | -0.07 | 0.84 | 20 |
| Tunisia | 2.28 | -0.07 | 0.87 | 24 |
| Mexico | 1.25 | -0.08 | 0.85 | 143 |
| Madagascar | 0.68 | -0.09 | 0.96 | 15 |
| Bosnia and Herzegovina | 3.84 | -0.12 | 0.69 | 11 |
| Oman | 4.69 | -0.12 | 0.92 | 17 |
| Hungary | 3.68 | -0.14 | 0.79 | 31 |
| Peru | 1.55 | -0.15 | 0.83 | 43 |
| Malta | 11.90 | -0.16 | 0.95 | 7 |
| Aruba | 18.81 | -0.17 | 1.00 | 0 |
| Colombia | 1.38 | -0.17 | 0.84 | 72 |
| Bermuda | 31.29 | -0.19 | 0.86 | 0 |
| Ecuador | 1.77 | -0.19 | 0.84 | 21 |
| Costa Rica | 3.57 | -0.20 | 0.90 | 16 |
| Sudan | 0.80 | -0.20 | 0.87 | 30 |
| Papua New Guinea | 1.41 | -0.21 | 0.82 | 14 |
| Haiti | 1.10 | -0.22 | 0.82 | 11 |
| Bulgaria | 3.25 | -0.23 | 0.81 | 22 |
| Albania | 3.39 | -0.26 | 0.72 | 6 |
| Suriname | 6.88 | -0.26 | 1.00 | 4 |
| Bahamas, The | 10.27 | -0.30 | 1.00 | 2 |
| Cabo Verde | 4.24 | -0.32 | 1.00 | 2 |
| Afghanistan | 0.53 | -0.33 | 0.72 | 15 |
| Uruguay | 3.76 | -0.33 | 0.83 | 12 |
| Nicaragua | 1.48 | -0.37 | 0.91 | 11 |
| Lesotho | 1.57 | -0.38 | 1.00 | 4 |
| Guinea-Bissau | 1.39 | -0.39 | 0.73 | 1 |
| Qatar | 7.06 | -0.39 | 0.86 | 23 |
| Maldives | 6.28 | -0.40 | 0.91 | 5 |
| Egypt, Arab Rep. | 0.76 | -0.42 | 0.86 | 83 |
| Timor-Leste | 2.06 | -0.43 | 0.80 | 3 |
| Kuwait | 4.44 | -0.44 | 0.98 | 16 |
| Myanmar | 0.62 | -0.44 | 0.78 | 40 |
| Argentina | 1.31 | -0.46 | 0.81 | 60 |
| Turks and Caicos Islands | 17.45 | -0.49 | 1.00 | 1 |
| Dominica | 9.28 | -0.50 | 1.00 | 1 |
| St. Lucia | 7.29 | -0.51 | 0.80 | 0 |
| Congo, Dem. Rep. | 0.25 | -0.52 | 0.86 | 18 |
| Ukraine | 0.96 | -0.52 | 0.59 | 37 |
| Iraq | 0.92 | -0.54 | 0.75 | 32 |
| Morocco | 0.79 | -0.55 | 0.88 | 28 |
| Somalia | 0.45 | -0.55 | 0.66 | 7 |
| Panama | 3.06 | -0.60 | 0.89 | 11 |
| Philippines | 0.55 | -0.60 | 0.81 | 53 |
| Luxembourg | 10.75 | -0.62 | 0.74 | 4 |
| Armenia | 2.25 | -0.65 | 0.77 | 2 |
| Bolivia | 1.13 | -0.66 | 0.70 | 13 |
| Russian Federation | 0.77 | -0.68 | 0.67 | 86 |
| Burundi | 0.35 | -0.72 | 0.75 | 4 |
| Bahrain | 4.47 | -0.72 | 0.85 | 6 |
| Mauritius | 3.42 | -0.77 | 0.76 | 3 |
| Honduras | 0.86 | -0.80 | 0.81 | 9 |
| Guyana | 2.98 | -0.82 | 1.00 | 2 |
| Trinidad and Tobago | 3.35 | -0.83 | 0.82 | 2 |
| Macao SAR, China | 8.85 | -0.85 | 0.89 | 2 |
| Antigua and Barbuda | 6.86 | -0.96 | 0.67 | 0 |
| Guatemala | 0.70 | -1.00 | 0.70 | 7 |
| Chad | 0.31 | -1.04 | 0.75 | 4 |
| Comoros | 1.18 | -1.06 | 0.75 | 1 |
| Dominican Republic | 1.06 | -1.09 | 0.83 | 12 |
| Gabon | 1.69 | -1.09 | 1.00 | 6 |
| Djibouti | 1.37 | -1.11 | 0.67 | 1 |
| Kyrgyz Republic | 0.67 | -1.15 | 0.59 | 4 |
| United Arab Emirates | 1.84 | -1.17 | 0.87 | 21 |
| Azerbaijan | 0.86 | -1.21 | 0.72 | 9 |
| El Salvador | 0.77 | -1.24 | 0.83 | 4 |
| Cayman Islands | 10.26 | -1.25 | 0.67 | 1 |
| Tajikistan | 0.43 | -1.29 | 0.71 | 2 |
| Mauritania | 0.66 | -1.30 | 0.82 | 3 |
| Curacao | 4.23 | -1.31 | 1.00 | 0 |
| Libya | 1.08 | -1.31 | 0.67 | 2 |
| Paraguay | 0.80 | -1.34 | 0.81 | 9 |
| Kazakhstan | 0.79 | -1.35 | 0.73 | 14 |
| Moldova | 1.13 | -1.36 | 0.56 | 3 |
| San Marino | 9.84 | -1.44 | 1.00 | 0 |
| Equatorial Guinea | 1.47 | -1.49 | 1.00 | 3 |
| Algeria | 0.33 | -1.54 | 0.90 | 12 |
| Cote d'Ivoire | 0.25 | -1.66 | 0.90 | 3 |
| Belarus | 0.57 | -1.78 | 0.67 | 4 |
| Angola | 0.17 | -2.09 | 0.67 | 4 |
| Uzbekistan | 0.12 | -2.45 | 0.80 | 1 |
| Turkmenistan | 0.17 | -3.06 | 0.50 | 0 |
| Sint Maarten (Dutch part) | 0.00 | -Inf |  | 0 |
| Andorra | 0.00 |  | 0.00 | 0 |
| American Samoa | 24.11 |  | 0.80 | 2 |
| Cuba | 1.09 |  | 0.70 | 12 |
| Eritrea | 1.24 |  | 0.86 | 5 |
| Faroe Islands | 27.39 |  | 1.00 | 4 |
| Gibraltar | 79.12 |  | 0.80 | 3 |
| Greenland | 213.43 |  | 0.90 | 12 |
| Guam | 17.93 |  | 0.75 | 1 |
| Isle of Man | 15.76 |  | 1.00 | 1 |
| Liechtenstein | 26.30 |  | 0.75 | 0 |
| St. Martin (French part) | 122.80 |  | 0.82 | 4 |
| Monaco | 68.43 |  | 1.00 | 2 |
| Northern Mariana Islands | 11.65 |  | 0.67 | 0 |
| New Caledonia | 7.37 |  | 1.00 | 1 |
| Korea, Dem. People's Rep. | 0.14 |  | 0.50 | 2 |
| South Sudan | 0.51 |  | 0.68 | 3 |
| Syrian Arab Republic | 0.80 |  | 0.62 | 16 |
| Taiwan, China |  |  | 0.90 | 146 |
| Venezuela, RB | 0.37 |  | 0.74 | 7 |
| British Virgin Islands | 44.40 |  | 1.00 | 1 |
| Virgin Islands (U.S.) | 90.62 |  | 0.85 | 6 |
| Yemen, Rep. | 0.47 |  | 0.82 | 17 |
| French Polynesia | 10.74 |  | 0.82 | 4 |

## 8.1 Machine Learning Approach

Add Xioayu’s description of the model.

## 8.2 Country Scores by Subject

Table. Country Scores on Academic Data Use by Subject of Article. Per capita measure. Year 2019.

| Country | Academic Data Use per capita - All | Academic Data Use per capita - Economics | Academic Data Use per capita - Psychology | Academic Data Use per capita - Sociology | data\_papers\_3yr\_Political Science\_pcap | Academic Data Use per capita - Medicine |
| --- | --- | --- | --- | --- | --- | --- |
| Greenland | 213.43 | 17.79 | 0.00 | 0.00 | 0.00 | 195.64 |
| St. Martin (French part) | 122.80 | 0.00 | 17.54 | 26.31 | 0.00 | 78.94 |
| Virgin Islands (U.S.) | 90.62 | 3.12 | 0.00 | 3.12 | 0.00 | 84.37 |
| Gibraltar | 79.12 | 0.00 | 0.00 | 0.00 | 0.00 | 79.12 |
| Palau | 74.07 | 0.00 | 0.00 | 0.00 | 0.00 | 74.07 |
| Monaco | 68.43 | 8.55 | 0.00 | 0.00 | 0.00 | 59.88 |
| Iceland | 59.17 | 4.62 | 3.70 | 1.85 | 5.55 | 42.53 |
| Tuvalu | 57.20 | 0.00 | 0.00 | 0.00 | 0.00 | 57.20 |
| Sao Tome and Principe | 49.60 | 1.55 | 0.00 | 1.55 | 0.00 | 46.50 |
| British Virgin Islands | 44.40 | 0.00 | 0.00 | 0.00 | 0.00 | 33.30 |
| New Zealand | 37.29 | 1.67 | 0.87 | 1.00 | 0.67 | 32.27 |
| St. Kitts and Nevis | 31.55 | 0.00 | 0.00 | 0.00 | 0.00 | 25.24 |
| Bermuda | 31.29 | 10.43 | 0.00 | 0.00 | 0.00 | 20.86 |
| Nauru | 30.97 | 0.00 | 30.97 | 0.00 | 0.00 | 0.00 |
| Seychelles | 30.73 | 3.41 | 0.00 | 0.00 | 0.00 | 27.32 |
| Faroe Islands | 27.39 | 0.00 | 0.00 | 0.00 | 0.00 | 27.39 |
| Liechtenstein | 26.30 | 8.77 | 0.00 | 0.00 | 0.00 | 8.77 |
| Norway | 24.56 | 2.74 | 1.06 | 0.81 | 0.87 | 18.26 |
| American Samoa | 24.11 | 0.00 | 0.00 | 6.03 | 0.00 | 18.08 |
| Ireland | 22.56 | 2.09 | 1.08 | 1.15 | 0.61 | 17.36 |
| Brunei Darussalam | 20.77 | 6.92 | 1.54 | 0.77 | 0.77 | 10.00 |
| Kiribati | 19.84 | 0.00 | 0.00 | 0.00 | 0.00 | 19.84 |
| Denmark | 19.38 | 1.83 | 0.34 | 0.63 | 0.34 | 15.54 |
| Aruba | 18.81 | 0.00 | 3.14 | 0.00 | 0.00 | 12.54 |
| Sweden | 18.13 | 1.82 | 0.68 | 0.91 | 0.78 | 13.30 |
| Guam | 17.93 | 1.99 | 0.00 | 1.99 | 0.00 | 13.95 |
| Grenada | 17.86 | 0.00 | 0.00 | 0.00 | 0.00 | 14.88 |
| Australia | 17.79 | 1.03 | 0.75 | 0.60 | 0.33 | 14.63 |
| Micronesia, Fed. Sts. | 17.57 | 0.00 | 0.00 | 0.00 | 0.00 | 17.57 |
| Turks and Caicos Islands | 17.45 | 8.73 | 0.00 | 0.00 | 0.00 | 8.73 |
| Finland | 17.21 | 2.17 | 0.72 | 1.03 | 0.54 | 12.07 |
| Singapore | 16.54 | 1.93 | 0.82 | 0.58 | 0.41 | 11.75 |
| Tonga | 15.95 | 0.00 | 0.00 | 0.00 | 0.00 | 15.95 |
| Isle of Man | 15.76 | 0.00 | 0.00 | 0.00 | 0.00 | 15.76 |
| St. Vincent and the Grenadines | 15.07 | 0.00 | 0.00 | 0.00 | 0.00 | 12.06 |
| Estonia | 14.82 | 2.01 | 0.75 | 0.50 | 1.26 | 8.29 |
| Vanuatu | 13.34 | 0.00 | 0.00 | 2.22 | 2.22 | 7.78 |
| Switzerland | 12.09 | 1.21 | 0.39 | 0.31 | 0.31 | 9.64 |
| Netherlands | 12.05 | 0.88 | 0.48 | 0.38 | 0.23 | 9.67 |
| Fiji | 11.99 | 0.75 | 0.00 | 1.12 | 0.00 | 9.36 |
| Belize | 11.96 | 0.00 | 0.85 | 0.85 | 0.00 | 9.39 |
| Malta | 11.90 | 1.98 | 0.66 | 1.32 | 0.00 | 7.27 |
| Samoa | 11.84 | 0.00 | 0.00 | 3.38 | 0.00 | 8.46 |
| Northern Mariana Islands | 11.65 | 5.83 | 0.00 | 0.00 | 0.00 | 5.83 |
| Marshall Islands | 11.34 | 0.00 | 0.00 | 0.00 | 0.00 | 11.34 |
| Hong Kong SAR, China | 11.23 | 1.07 | 1.24 | 0.36 | 0.18 | 7.64 |
| Cyprus | 11.12 | 0.56 | 2.50 | 1.39 | 0.56 | 6.12 |
| Luxembourg | 10.75 | 4.30 | 1.08 | 1.61 | 0.00 | 2.15 |
| French Polynesia | 10.74 | 0.00 | 0.00 | 0.00 | 0.00 | 10.74 |
| Slovenia | 10.53 | 1.76 | 0.64 | 0.32 | 0.64 | 5.91 |
| Barbados | 10.45 | 1.16 | 1.16 | 0.00 | 0.00 | 6.97 |
| Georgia | 10.39 | 0.72 | 0.36 | 0.27 | 0.09 | 8.42 |
| Bahamas, The | 10.27 | 1.71 | 0.00 | 0.00 | 0.00 | 7.70 |
| Cayman Islands | 10.26 | 0.00 | 0.00 | 0.00 | 0.00 | 5.13 |
| Bhutan | 10.05 | 1.75 | 0.44 | 0.00 | 0.44 | 6.55 |
| San Marino | 9.84 | 0.00 | 0.00 | 0.00 | 0.00 | 9.84 |
| Montenegro | 9.65 | 4.29 | 0.54 | 0.00 | 0.54 | 3.75 |
| Canada | 9.59 | 0.68 | 0.35 | 0.35 | 0.25 | 7.76 |
| Dominica | 9.28 | 0.00 | 0.00 | 0.00 | 0.00 | 4.64 |
| Botswana | 9.26 | 0.72 | 1.16 | 0.29 | 0.29 | 5.93 |
| Macao SAR, China | 8.85 | 2.08 | 0.52 | 0.00 | 0.00 | 4.68 |
| Croatia | 8.36 | 1.39 | 0.49 | 0.41 | 0.49 | 4.18 |
| Latvia | 8.01 | 0.87 | 0.17 | 0.17 | 0.70 | 3.14 |
| Lithuania | 7.64 | 1.55 | 0.24 | 0.24 | 0.60 | 3.46 |
| Israel | 7.55 | 0.66 | 0.29 | 0.29 | 0.18 | 6.07 |
| Malaysia | 7.44 | 0.85 | 1.20 | 0.43 | 0.22 | 3.14 |
| Greece | 7.37 | 0.99 | 0.59 | 0.31 | 0.44 | 4.66 |
| New Caledonia | 7.37 | 0.00 | 0.00 | 0.00 | 2.46 | 3.69 |
| St. Lucia | 7.29 | 3.65 | 0.00 | 0.00 | 0.00 | 1.82 |
| Qatar | 7.06 | 0.94 | 0.00 | 0.12 | 0.00 | 5.77 |
| Portugal | 7.00 | 1.36 | 0.32 | 0.42 | 0.16 | 4.15 |
| Solomon Islands | 6.97 | 1.99 | 0.00 | 1.00 | 0.00 | 3.98 |
| Congo, Rep. | 6.88 | 0.19 | 0.00 | 0.00 | 0.06 | 6.20 |
| Suriname | 6.88 | 0.00 | 0.57 | 0.57 | 0.00 | 5.73 |
| Antigua and Barbuda | 6.86 | 0.00 | 0.00 | 0.00 | 0.00 | 3.43 |
| Austria | 6.49 | 0.86 | 0.19 | 0.19 | 0.30 | 4.73 |
| Mongolia | 6.41 | 0.83 | 0.10 | 0.00 | 0.00 | 5.06 |
| Maldives | 6.28 | 2.51 | 0.00 | 0.00 | 0.00 | 3.14 |
| Puerto Rico | 6.26 | 0.21 | 0.31 | 0.10 | 0.00 | 5.64 |
| Kosovo | 5.96 | 2.24 | 0.37 | 0.00 | 0.56 | 1.68 |
| Belgium | 5.74 | 0.55 | 0.06 | 0.23 | 0.20 | 4.50 |
| North Macedonia | 5.62 | 1.77 | 0.00 | 0.00 | 0.48 | 2.73 |
| Eswatini | 5.52 | 0.00 | 0.58 | 0.00 | 0.00 | 4.65 |
| Korea, Rep. | 5.51 | 0.38 | 0.10 | 0.05 | 0.03 | 4.79 |
| Slovak Republic | 5.44 | 1.28 | 0.18 | 0.06 | 0.31 | 3.06 |
| Czech Republic | 4.99 | 1.44 | 0.22 | 0.13 | 0.09 | 2.54 |
| Czechia | 4.97 | 1.44 | 0.22 | 0.12 | 0.09 | 2.53 |
| Jordan | 4.88 | 0.53 | 0.49 | 0.16 | 0.07 | 3.10 |
| Namibia | 4.81 | 0.27 | 0.53 | 0.40 | 0.27 | 2.67 |
| Spain | 4.79 | 0.64 | 0.14 | 0.18 | 0.06 | 3.65 |
| Oman | 4.69 | 0.47 | 0.27 | 0.13 | 0.13 | 2.95 |
| South Africa | 4.58 | 0.55 | 0.29 | 0.18 | 0.12 | 3.14 |
| Bahrain | 4.47 | 1.02 | 0.20 | 0.00 | 0.20 | 2.23 |
| Kuwait | 4.44 | 0.63 | 0.55 | 0.08 | 0.16 | 2.69 |
| Italy | 4.40 | 0.56 | 0.13 | 0.14 | 0.08 | 3.31 |
| Guinea | 4.38 | 0.13 | 0.05 | 0.03 | 0.08 | 4.07 |
| Cabo Verde | 4.24 | 1.21 | 0.00 | 0.00 | 0.00 | 3.03 |
| Curacao | 4.23 | 0.00 | 0.00 | 0.00 | 0.00 | 4.23 |
| Serbia | 4.22 | 1.01 | 0.24 | 0.00 | 0.24 | 2.40 |
| Germany | 4.01 | 0.58 | 0.09 | 0.15 | 0.12 | 2.90 |
| Iran, Islamic Rep. | 3.98 | 0.17 | 0.32 | 0.02 | 0.05 | 3.32 |
| West Bank and Gaza | 3.98 | 0.21 | 0.36 | 0.43 | 0.43 | 2.42 |
| Ghana | 3.85 | 0.66 | 0.21 | 0.20 | 0.08 | 2.26 |
| Bosnia and Herzegovina | 3.84 | 0.81 | 0.30 | 0.50 | 0.40 | 1.62 |
| Saudi Arabia | 3.84 | 0.18 | 0.27 | 0.02 | 0.06 | 3.18 |
| United Kingdom | 3.83 | 0.31 | 0.15 | 0.20 | 0.10 | 2.99 |
| Uruguay | 3.76 | 0.19 | 0.29 | 0.10 | 0.00 | 2.99 |
| Jamaica | 3.73 | 0.23 | 0.11 | 0.23 | 0.23 | 2.83 |
| Hungary | 3.68 | 0.82 | 0.07 | 0.14 | 0.14 | 2.22 |
| Chile | 3.57 | 0.67 | 0.11 | 0.05 | 0.14 | 2.52 |
| Costa Rica | 3.57 | 0.59 | 0.26 | 0.13 | 0.00 | 2.44 |
| Mauritius | 3.42 | 1.05 | 0.00 | 0.00 | 0.00 | 0.79 |
| Albania | 3.39 | 1.05 | 0.35 | 0.00 | 0.23 | 1.28 |
| Japan | 3.37 | 0.24 | 0.06 | 0.04 | 0.03 | 2.89 |
| Poland | 3.37 | 0.61 | 0.10 | 0.11 | 0.13 | 2.05 |
| Trinidad and Tobago | 3.35 | 0.00 | 0.24 | 0.24 | 0.00 | 2.63 |
| Romania | 3.27 | 0.65 | 0.09 | 0.21 | 0.22 | 1.55 |
| Bulgaria | 3.25 | 0.67 | 0.19 | 0.19 | 0.33 | 1.58 |
| Nepal | 3.22 | 0.31 | 0.03 | 0.08 | 0.09 | 2.54 |
| Lebanon | 3.21 | 0.15 | 0.05 | 0.10 | 0.10 | 2.67 |
| Malawi | 3.08 | 0.29 | 0.02 | 0.02 | 0.07 | 2.47 |
| Panama | 3.06 | 0.08 | 0.00 | 0.00 | 0.00 | 2.98 |
| Guyana | 2.98 | 0.43 | 0.00 | 0.43 | 0.00 | 2.13 |
| Kenya | 2.94 | 0.32 | 0.16 | 0.12 | 0.08 | 1.71 |
| France | 2.85 | 0.32 | 0.02 | 0.07 | 0.06 | 2.26 |
| Gambia, The | 2.84 | 0.28 | 0.00 | 0.00 | 0.00 | 2.56 |
| Sri Lanka | 2.71 | 0.35 | 0.05 | 0.05 | 0.05 | 1.88 |
| United States | 2.70 | 0.19 | 0.09 | 0.08 | 0.06 | 2.19 |
| Lao PDR | 2.65 | 0.65 | 0.09 | 0.05 | 0.14 | 1.67 |
| Zimbabwe | 2.46 | 0.18 | 0.09 | 0.14 | 0.16 | 1.59 |
| Turkey | 2.36 | 0.25 | 0.49 | 0.15 | 0.09 | 1.23 |
| Turkiye | 2.33 | 0.25 | 0.48 | 0.15 | 0.09 | 1.21 |
| Tunisia | 2.28 | 0.37 | 0.00 | 0.09 | 0.06 | 1.65 |
| Armenia | 2.25 | 0.79 | 0.11 | 0.00 | 0.11 | 1.01 |
| Uganda | 2.23 | 0.17 | 0.05 | 0.05 | 0.06 | 1.75 |
| Thailand | 2.13 | 0.24 | 0.20 | 0.05 | 0.05 | 1.39 |
| Rwanda | 2.06 | 0.21 | 0.05 | 0.05 | 0.18 | 1.29 |
| Timor-Leste | 2.06 | 0.26 | 0.00 | 0.52 | 0.00 | 1.29 |
| Liberia | 2.03 | 0.00 | 0.00 | 0.14 | 0.20 | 1.62 |
| Sierra Leone | 1.96 | 0.17 | 0.04 | 0.17 | 0.09 | 1.41 |
| United Arab Emirates | 1.84 | 0.38 | 0.00 | 0.10 | 0.07 | 1.16 |
| Brazil | 1.84 | 0.16 | 0.05 | 0.03 | 0.04 | 1.50 |
| Zambia | 1.83 | 0.24 | 0.02 | 0.09 | 0.04 | 1.33 |
| Ecuador | 1.77 | 0.17 | 0.13 | 0.12 | 0.06 | 1.23 |
| Gabon | 1.69 | 0.00 | 0.00 | 0.00 | 0.00 | 1.53 |
| Niger | 1.63 | 0.10 | 0.03 | 0.03 | 0.00 | 1.39 |
| Lesotho | 1.57 | 0.16 | 0.00 | 0.16 | 0.16 | 0.94 |
| Ethiopia | 1.55 | 0.15 | 0.05 | 0.03 | 0.03 | 1.19 |
| Peru | 1.55 | 0.21 | 0.07 | 0.02 | 0.02 | 1.19 |
| Tanzania | 1.54 | 0.24 | 0.03 | 0.04 | 0.04 | 1.03 |
| Cambodia | 1.52 | 0.26 | 0.02 | 0.04 | 0.04 | 1.09 |
| Nicaragua | 1.48 | 0.15 | 0.05 | 0.00 | 0.00 | 1.22 |
| Togo | 1.48 | 0.16 | 0.04 | 0.00 | 0.00 | 1.24 |
| Equatorial Guinea | 1.47 | 0.25 | 0.00 | 0.00 | 0.25 | 0.98 |
| Cameroon | 1.44 | 0.06 | 0.05 | 0.03 | 0.04 | 1.20 |
| Papua New Guinea | 1.41 | 0.08 | 0.00 | 0.04 | 0.04 | 1.22 |
| Guinea-Bissau | 1.39 | 0.17 | 0.00 | 0.00 | 0.00 | 1.21 |
| Colombia | 1.38 | 0.19 | 0.03 | 0.06 | 0.06 | 0.99 |
| Djibouti | 1.37 | 0.00 | 0.34 | 0.00 | 0.00 | 1.03 |
| Senegal | 1.37 | 0.14 | 0.00 | 0.02 | 0.04 | 1.08 |
| Benin | 1.36 | 0.11 | 0.08 | 0.00 | 0.00 | 1.05 |
| Nigeria | 1.35 | 0.17 | 0.11 | 0.05 | 0.05 | 0.72 |
| Argentina | 1.31 | 0.13 | 0.02 | 0.01 | 0.01 | 1.12 |
| Burkina Faso | 1.28 | 0.07 | 0.02 | 0.03 | 0.03 | 1.07 |
| Indonesia | 1.28 | 0.16 | 0.23 | 0.06 | 0.05 | 0.39 |
| Mexico | 1.25 | 0.15 | 0.03 | 0.03 | 0.01 | 0.98 |
| Eritrea | 1.24 | 0.00 | 0.00 | 0.00 | 0.00 | 1.24 |
| Comoros | 1.18 | 0.00 | 0.00 | 0.00 | 0.00 | 1.18 |
| Bolivia | 1.13 | 0.12 | 0.03 | 0.03 | 0.06 | 0.87 |
| Moldova | 1.13 | 0.00 | 0.13 | 0.13 | 0.00 | 0.75 |
| Central African Republic | 1.12 | 0.07 | 0.00 | 0.00 | 0.07 | 0.98 |
| Haiti | 1.10 | 0.00 | 0.00 | 0.00 | 0.06 | 1.01 |
| Cuba | 1.09 | 0.00 | 0.03 | 0.09 | 0.03 | 0.88 |
| Libya | 1.08 | 0.15 | 0.00 | 0.00 | 0.10 | 0.74 |
| Vietnam | 1.08 | 0.27 | 0.04 | 0.01 | 0.03 | 0.64 |
| Dominican Republic | 1.06 | 0.09 | 0.00 | 0.00 | 0.03 | 0.90 |
| China | 0.99 | 0.13 | 0.01 | 0.01 | 0.01 | 0.79 |
| Ukraine | 0.96 | 0.15 | 0.07 | 0.05 | 0.11 | 0.32 |
| Bangladesh | 0.94 | 0.15 | 0.02 | 0.01 | 0.02 | 0.66 |
| Iraq | 0.92 | 0.05 | 0.05 | 0.07 | 0.06 | 0.65 |
| Mali | 0.92 | 0.14 | 0.00 | 0.03 | 0.03 | 0.63 |
| Azerbaijan | 0.86 | 0.17 | 0.07 | 0.00 | 0.00 | 0.60 |
| Honduras | 0.86 | 0.07 | 0.00 | 0.00 | 0.07 | 0.68 |
| Pakistan | 0.81 | 0.11 | 0.07 | 0.02 | 0.03 | 0.50 |
| Paraguay | 0.80 | 0.05 | 0.00 | 0.00 | 0.00 | 0.71 |
| Sudan | 0.80 | 0.03 | 0.03 | 0.02 | 0.02 | 0.69 |
| Syrian Arab Republic | 0.80 | 0.08 | 0.06 | 0.06 | 0.16 | 0.43 |
| Kazakhstan | 0.79 | 0.20 | 0.04 | 0.00 | 0.07 | 0.45 |
| Morocco | 0.79 | 0.07 | 0.03 | 0.03 | 0.02 | 0.55 |
| Russian Federation | 0.77 | 0.17 | 0.05 | 0.03 | 0.08 | 0.36 |
| El Salvador | 0.77 | 0.10 | 0.00 | 0.05 | 0.05 | 0.57 |
| Egypt, Arab Rep. | 0.76 | 0.08 | 0.01 | 0.01 | 0.02 | 0.59 |
| Mozambique | 0.74 | 0.10 | 0.05 | 0.00 | 0.02 | 0.54 |
| Guatemala | 0.70 | 0.06 | 0.02 | 0.02 | 0.02 | 0.58 |
| Madagascar | 0.68 | 0.09 | 0.00 | 0.01 | 0.00 | 0.57 |
| Kyrgyz Republic | 0.67 | 0.15 | 0.05 | 0.05 | 0.00 | 0.36 |
| Mauritania | 0.66 | 0.15 | 0.00 | 0.00 | 0.00 | 0.44 |
| Myanmar | 0.62 | 0.07 | 0.02 | 0.02 | 0.05 | 0.41 |
| Belarus | 0.57 | 0.14 | 0.04 | 0.04 | 0.04 | 0.32 |
| Philippines | 0.55 | 0.10 | 0.04 | 0.03 | 0.02 | 0.29 |
| Afghanistan | 0.53 | 0.05 | 0.01 | 0.03 | 0.05 | 0.38 |
| South Sudan | 0.51 | 0.00 | 0.06 | 0.03 | 0.06 | 0.36 |
| Yemen, Rep. | 0.47 | 0.03 | 0.02 | 0.01 | 0.03 | 0.33 |
| Somalia | 0.45 | 0.06 | 0.06 | 0.00 | 0.09 | 0.22 |
| Tajikistan | 0.43 | 0.14 | 0.04 | 0.00 | 0.04 | 0.18 |
| India | 0.39 | 0.04 | 0.01 | 0.01 | 0.01 | 0.29 |
| Venezuela, RB | 0.37 | 0.07 | 0.00 | 0.01 | 0.01 | 0.28 |
| Burundi | 0.35 | 0.03 | 0.00 | 0.00 | 0.03 | 0.23 |
| Algeria | 0.33 | 0.02 | 0.02 | 0.00 | 0.01 | 0.29 |
| Chad | 0.31 | 0.02 | 0.00 | 0.00 | 0.04 | 0.23 |
| Cote d'Ivoire | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 |
| Congo, Dem. Rep. | 0.25 | 0.01 | 0.00 | 0.00 | 0.00 | 0.20 |
| Angola | 0.17 | 0.02 | 0.00 | 0.00 | 0.00 | 0.14 |
| Turkmenistan | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 |
| Korea, Dem. People's Rep. | 0.14 | 0.03 | 0.00 | 0.01 | 0.03 | 0.06 |
| Uzbekistan | 0.12 | 0.03 | 0.00 | 0.00 | 0.01 | 0.08 |
| Andorra | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Sint Maarten (Dutch part) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Taiwan, China |  |  |  |  |  |  |

Table. Country Scores on Academic Data Use by Subject of Article. Regression Adjusted measure. Year 2019.

| Country | Academic Data Use - Adjusted - All | Academic Data Use - Adjusted - Economics | Academic Data Use - Adjusted - Psychology | Academic Data Use - Adjusted - Sociology | Academic Data Use - Adjusted - Political Science | Academic Data Use - Adjusted - Medicine |
| --- | --- | --- | --- | --- | --- | --- |
| Sao Tome and Principe | 2.08 | 0.47 | -Inf | 0.58 | -Inf | 2.33 |
| New Zealand | 1.81 | 0.64 | 0.76 | 1.18 | 0.80 | 2.08 |
| Australia | 1.58 | 0.78 | 1.47 | 1.60 | 0.99 | 1.83 |
| Malawi | 1.31 | 1.21 | -0.72 | -0.88 | 0.59 | 1.45 |
| Guinea | 1.29 | -0.01 | -0.12 | -0.91 | 0.27 | 1.59 |
| Congo, Rep. | 1.26 | -0.24 | -Inf | -Inf | -0.62 | 1.52 |
| Sweden | 1.25 | 0.92 | 0.83 | 1.44 | 1.30 | 1.37 |
| Kiribati | 1.24 | -Inf | -Inf | -Inf | -Inf | 1.54 |
| Norway | 1.23 | 0.95 | 0.80 | 0.87 | 0.96 | 1.36 |
| Iceland | 1.22 | 0.37 | 0.59 | 0.13 | 1.28 | 1.27 |
| Kenya | 1.14 | 1.21 | 1.52 | 1.27 | 0.85 | 1.00 |
| Tuvalu | 1.14 | -Inf | -Inf | -Inf | -Inf | 1.41 |
| Uganda | 1.13 | 0.85 | 0.67 | 0.44 | 0.79 | 1.28 |
| Ghana | 1.12 | 1.57 | 1.37 | 1.36 | 0.46 | 0.98 |
| Canada | 1.11 | 0.54 | 0.94 | 1.30 | 0.94 | 1.34 |
| South Africa | 1.11 | 1.19 | 1.63 | 1.35 | 0.95 | 1.16 |
| Ethiopia | 1.09 | 1.15 | 1.09 | 0.57 | 0.71 | 1.22 |
| Denmark | 1.08 | 0.64 | -0.22 | 0.71 | 0.12 | 1.28 |
| Nepal | 1.07 | 0.99 | -0.28 | 0.56 | 0.75 | 1.23 |
| Iran, Islamic Rep. | 1.05 | 0.12 | 1.88 | -0.68 | 0.21 | 1.30 |
| Malaysia | 1.05 | 0.98 | 2.34 | 1.58 | 0.92 | 0.61 |
| Finland | 1.02 | 0.88 | 0.58 | 1.22 | 0.61 | 1.08 |
| Netherlands | 1.00 | 0.39 | 0.74 | 0.87 | 0.36 | 1.21 |
| Vanuatu | 1.00 | -Inf | -Inf | 1.23 | 1.36 | 0.77 |
| Ireland | 0.99 | 0.49 | 0.63 | 1.06 | 0.44 | 1.15 |
| Palau | 0.93 | -Inf | -Inf | -Inf | -Inf | 1.24 |
| Georgia | 0.91 | 0.22 | 0.24 | 0.06 | -0.98 | 1.09 |
| Micronesia, Fed. Sts. | 0.88 | -Inf | -Inf | -Inf | -Inf | 1.19 |
| Niger | 0.86 | 0.38 | -0.01 | -0.20 | -Inf | 1.06 |
| Zimbabwe | 0.81 | 0.44 | 0.57 | 0.86 | 1.10 | 0.75 |
| Nigeria | 0.77 | 1.09 | 1.80 | 1.23 | 1.19 | 0.57 |
| Tanzania | 0.77 | 1.21 | 0.30 | 0.41 | 0.47 | 0.76 |
| Korea, Rep. | 0.73 | 0.17 | -0.11 | -0.52 | -1.09 | 1.03 |
| Solomon Islands | 0.70 | 1.42 | -Inf | 0.96 | -Inf | 0.47 |
| Jordan | 0.69 | 0.56 | 1.32 | 0.31 | -0.56 | 0.63 |
| Singapore | 0.67 | 0.40 | 0.37 | 0.43 | 0.08 | 0.76 |
| Switzerland | 0.65 | 0.28 | 0.02 | 0.16 | 0.16 | 0.85 |
| China | 0.63 | 1.03 | 0.38 | 0.40 | 0.36 | 0.88 |
| Greece | 0.63 | 0.64 | 0.99 | 0.60 | 0.96 | 0.59 |
| Brazil | 0.62 | 0.47 | 0.47 | 0.43 | 0.63 | 0.86 |
| Hong Kong SAR, China | 0.61 | 0.19 | 1.19 | 0.27 | -0.41 | 0.65 |
| Belize | 0.60 | -Inf | 0.20 | 0.14 | -Inf | 0.70 |
| Rwanda | 0.60 | 0.54 | -0.04 | -0.16 | 1.17 | 0.50 |
| Botswana | 0.59 | -0.02 | 1.09 | -0.18 | -0.12 | 0.52 |
| Fiji | 0.59 | -0.30 | -Inf | 0.67 | -Inf | 0.71 |
| Spain | 0.58 | 0.67 | 0.24 | 0.80 | -0.22 | 0.75 |
| Japan | 0.56 | 0.13 | -0.06 | -0.11 | -0.28 | 0.87 |
| Italy | 0.55 | 0.62 | 0.30 | 0.69 | 0.17 | 0.72 |
| Sierra Leone | 0.51 | 0.26 | -0.39 | 0.82 | 0.22 | 0.53 |
| Estonia | 0.50 | 0.34 | -0.04 | -0.24 | 0.73 | 0.31 |
| Portugal | 0.49 | 0.85 | 0.29 | 0.82 | -0.12 | 0.38 |
| United States | 0.49 | 0.09 | 0.69 | 1.07 | 0.70 | 0.77 |
| Germany | 0.47 | 0.67 | 0.00 | 0.90 | 0.61 | 0.60 |
| Tonga | 0.47 | -Inf | -Inf | -Inf | -Inf | 0.79 |
| Mongolia | 0.46 | 0.40 | -1.00 | -Inf | -Inf | 0.61 |
| Israel | 0.45 | 0.00 | 0.05 | 0.33 | -0.12 | 0.65 |
| Indonesia | 0.44 | 0.72 | 2.33 | 1.25 | 1.03 | -0.30 |
| Liberia | 0.44 | -Inf | -Inf | 0.37 | 0.87 | 0.57 |
| West Bank and Gaza | 0.44 | -0.42 | 0.79 | 0.97 | 1.04 | 0.31 |
| Bhutan | 0.43 | 0.56 | -0.34 | -Inf | -0.23 | 0.36 |
| Croatia | 0.42 | 0.57 | 0.27 | 0.30 | 0.52 | 0.13 |
| United Kingdom | 0.41 | 0.02 | 0.42 | 1.08 | 0.37 | 0.61 |
| Seychelles | 0.41 | -0.14 | -Inf | -Inf | -Inf | 0.63 |
| Zambia | 0.40 | 0.59 | -1.11 | 0.46 | -0.39 | 0.46 |
| Samoa | 0.38 | -Inf | -Inf | 1.13 | -Inf | 0.37 |
| Gambia, The | 0.33 | 0.11 | -Inf | -Inf | -Inf | 0.57 |
| Burkina Faso | 0.30 | -0.41 | -0.93 | -0.34 | -0.27 | 0.49 |
| Bangladesh | 0.30 | 0.81 | 0.10 | -0.60 | 0.21 | 0.36 |
| Slovenia | 0.29 | 0.36 | 0.01 | -0.45 | 0.29 | 0.10 |
| Pakistan | 0.27 | 0.69 | 1.47 | 0.26 | 0.54 | 0.22 |
| Thailand | 0.27 | 0.30 | 1.19 | 0.01 | 0.11 | 0.28 |
| Cameroon | 0.26 | -0.61 | 0.08 | -0.63 | -0.17 | 0.46 |
| Sri Lanka | 0.25 | 0.35 | -0.75 | -0.60 | -0.57 | 0.30 |
| Poland | 0.24 | 0.63 | -0.16 | 0.23 | 0.46 | 0.18 |
| St. Kitts and Nevis | 0.23 | -Inf | -Inf | -Inf | -Inf | 0.34 |
| Turkey | 0.23 | 0.17 | 1.97 | 1.10 | 0.62 | 0.03 |
| Kosovo | 0.23 | 1.19 | 0.00 | -Inf | 0.53 | -0.67 |
| Turkiye | 0.22 | 0.17 | 1.97 | 1.10 | 0.62 | 0.02 |
| Brunei Darussalam | 0.20 | 0.81 | -0.22 | -0.67 | -0.62 | -0.15 |
| Namibia | 0.20 | -0.71 | 0.62 | 0.38 | 0.04 | -0.02 |
| Chile | 0.19 | 0.59 | -0.32 | -0.77 | 0.23 | 0.26 |
| Saudi Arabia | 0.18 | -0.84 | 0.65 | -1.64 | -0.55 | 0.43 |
| Grenada | 0.17 | -Inf | -Inf | -Inf | -Inf | 0.32 |
| Austria | 0.15 | 0.09 | -0.57 | -0.24 | 0.24 | 0.26 |
| Belgium | 0.15 | -0.21 | -1.56 | 0.15 | 0.03 | 0.33 |
| Marshall Islands | 0.14 | -Inf | -Inf | -Inf | -Inf | 0.44 |
| Serbia | 0.14 | 0.73 | 0.09 | -Inf | 0.29 | -0.02 |
| Togo | 0.14 | 0.13 | -0.52 | -Inf | -Inf | 0.32 |
| France | 0.13 | 0.06 | -1.36 | 0.11 | -0.05 | 0.35 |
| Mozambique | 0.13 | 0.45 | 0.77 | -Inf | -0.24 | 0.19 |
| Cyprus | 0.12 | -1.06 | 1.04 | 0.67 | -0.20 | -0.09 |
| Eswatini | 0.12 | -Inf | 0.33 | -Inf | -Inf | 0.30 |
| Czech Republic | 0.10 | 0.85 | -0.15 | -0.42 | -0.69 | -0.15 |
| St. Vincent and the Grenadines | 0.10 | -Inf | -Inf | -Inf | -Inf | 0.20 |
| Czechia | 0.09 | 0.84 | -0.16 | -0.43 | -0.70 | -0.17 |
| India | 0.09 | 0.37 | 0.63 | 0.20 | 0.44 | 0.25 |
| Lithuania | 0.09 | 0.39 | -0.78 | -0.55 | 0.41 | -0.30 |
| Latvia | 0.09 | -0.25 | -1.22 | -1.02 | 0.41 | -0.46 |
| Senegal | 0.09 | 0.04 | -Inf | -1.10 | -0.34 | 0.24 |
| Cambodia | 0.08 | 0.51 | -1.19 | -0.51 | -0.45 | 0.13 |
| Lao PDR | 0.07 | 0.75 | -0.42 | -1.07 | 0.08 | 0.00 |
| Slovak Republic | 0.06 | 0.57 | -0.59 | -1.46 | 0.18 | -0.11 |
| North Macedonia | 0.04 | 0.81 | -Inf | -Inf | 0.32 | -0.30 |
| Montenegro | 0.04 | 1.05 | -0.55 | -Inf | -0.37 | -0.53 |
| Central African Republic | 0.04 | -0.53 | -Inf | -Inf | -0.04 | 0.24 |
| Romania | 0.02 | 0.48 | -0.60 | 0.54 | 0.63 | -0.30 |
| Barbados | 0.00 | -0.42 | -0.05 | -Inf | -Inf | -0.06 |
| Jamaica | 0.00 | -0.80 | -0.84 | -0.10 | -0.03 | 0.10 |
| Nauru | -0.01 | -Inf | 1.45 | -Inf | -Inf | -Inf |
| Benin | -0.02 | -0.32 | 0.20 | -Inf | -Inf | 0.09 |
| Vietnam | -0.02 | 0.86 | 0.19 | -0.82 | -0.11 | -0.12 |
| Lebanon | -0.03 | -1.09 | -1.40 | -0.58 | -0.53 | 0.18 |
| Puerto Rico | -0.03 | -1.52 | -0.40 | -1.27 | -Inf | 0.26 |
| Mali | -0.07 | 0.27 | -Inf | -0.35 | -0.28 | -0.08 |
| Tunisia | -0.07 | 0.20 | -Inf | -0.30 | -0.66 | 0.00 |
| Mexico | -0.08 | 0.03 | -0.34 | -0.07 | -0.95 | 0.12 |
| Madagascar | -0.09 | 0.15 | -Inf | -1.05 | -Inf | 0.10 |
| Bosnia and Herzegovina | -0.12 | 0.29 | 0.00 | 0.63 | 0.46 | -0.60 |
| Oman | -0.12 | -0.47 | -0.25 | -0.72 | -0.69 | -0.17 |
| Hungary | -0.14 | 0.36 | -1.26 | -0.31 | -0.29 | -0.23 |
| Peru | -0.15 | -0.01 | -0.07 | -1.16 | -1.14 | 0.00 |
| Malta | -0.16 | -0.21 | -0.83 | 0.07 | -Inf | -0.28 |
| Aruba | -0.17 | -Inf | -0.06 | -Inf | -Inf | -0.23 |
| Colombia | -0.17 | 0.01 | -0.66 | 0.12 | 0.14 | -0.08 |
| Bermuda | -0.19 | 0.25 | -Inf | -Inf | -Inf | -0.24 |
| Ecuador | -0.19 | -0.39 | 0.26 | 0.23 | -0.42 | -0.15 |
| Costa Rica | -0.20 | -0.01 | -0.06 | -0.58 | -Inf | -0.18 |
| Sudan | -0.20 | -1.18 | -0.19 | -0.88 | -0.42 | 0.04 |
| Papua New Guinea | -0.21 | -0.99 | -Inf | -0.94 | -0.87 | 0.02 |
| Haiti | -0.22 | -Inf | -Inf | -Inf | -0.15 | 0.07 |
| Bulgaria | -0.23 | 0.19 | -0.25 | -0.06 | 0.54 | -0.55 |
| Albania | -0.26 | 0.54 | 0.11 | -Inf | -0.14 | -0.85 |
| Suriname | -0.26 | -Inf | -0.46 | -0.37 | -Inf | -0.08 |
| Bahamas, The | -0.30 | -0.35 | -Inf | -Inf | -Inf | -0.22 |
| Cabo Verde | -0.32 | 0.32 | -Inf | -Inf | -Inf | -0.31 |
| Afghanistan | -0.33 | -0.33 | -1.18 | -0.17 | 0.58 | -0.29 |
| Uruguay | -0.33 | -1.35 | -0.23 | -1.16 | -Inf | -0.16 |
| Nicaragua | -0.37 | -0.54 | -0.89 | -Inf | -Inf | -0.19 |
| Lesotho | -0.38 | -0.62 | -Inf | -0.20 | -0.10 | -0.55 |
| Guinea-Bissau | -0.39 | -0.39 | -Inf | -Inf | -Inf | -0.19 |
| Qatar | -0.39 | -0.57 | -Inf | -1.56 | -Inf | -0.17 |
| Maldives | -0.40 | 0.50 | -Inf | -Inf | -Inf | -0.73 |
| Egypt, Arab Rep. | -0.42 | -0.38 | -1.08 | -0.89 | -0.66 | -0.25 |
| Timor-Leste | -0.43 | -0.50 | -Inf | 0.59 | -Inf | -0.55 |
| Kuwait | -0.44 | -0.49 | 0.15 | -1.51 | -0.79 | -0.53 |
| Myanmar | -0.44 | -0.37 | -0.64 | -0.32 | 0.41 | -0.45 |
| Argentina | -0.46 | -0.59 | -1.33 | -2.18 | -2.17 | -0.18 |
| Turks and Caicos Islands | -0.49 | 0.39 | -Inf | -Inf | -Inf | -0.85 |
| Dominica | -0.50 | -Inf | -Inf | -Inf | -Inf | -0.87 |
| St. Lucia | -0.51 | 0.54 | -Inf | -Inf | -Inf | -1.55 |
| Congo, Dem. Rep. | -0.52 | -1.16 | -Inf | -Inf | -1.31 | -0.33 |
| Ukraine | -0.52 | -0.19 | 0.05 | -0.19 | 0.68 | -1.19 |
| Iraq | -0.54 | -1.24 | -0.23 | 0.21 | 0.10 | -0.46 |
| Morocco | -0.55 | -0.71 | -0.70 | -0.62 | -0.98 | -0.50 |
| Somalia | -0.55 | -0.22 | 0.59 | -Inf | 0.76 | -0.94 |
| Panama | -0.60 | -2.32 | -Inf | -Inf | -Inf | -0.22 |
| Philippines | -0.60 | 0.03 | 0.28 | 0.09 | -0.11 | -0.82 |
| Luxembourg | -0.62 | 0.15 | -0.69 | 0.06 | -Inf | -1.83 |
| Armenia | -0.65 | 0.27 | -1.00 | -Inf | -0.85 | -1.07 |
| Bolivia | -0.66 | -0.82 | -1.37 | -1.30 | -0.56 | -0.53 |
| Russian Federation | -0.68 | 0.02 | 0.00 | -0.16 | 0.79 | -0.99 |
| Burundi | -0.72 | -0.91 | -Inf | -Inf | -0.33 | -0.78 |
| Bahrain | -0.72 | -0.37 | -1.34 | -Inf | -1.06 | -1.02 |
| Mauritius | -0.77 | -0.08 | -Inf | -Inf | -Inf | -1.86 |
| Honduras | -0.80 | -1.20 | -Inf | -Inf | -0.33 | -0.64 |
| Guyana | -0.82 | -0.90 | -Inf | -0.36 | -Inf | -0.80 |
| Trinidad and Tobago | -0.83 | -Inf | -1.01 | -0.84 | -Inf | -0.69 |
| Macao SAR, China | -0.85 | -0.61 | -1.44 | -Inf | -Inf | -1.08 |
| Antigua and Barbuda | -0.96 | -Inf | -Inf | -Inf | -Inf | -1.31 |
| Guatemala | -1.00 | -1.32 | -1.53 | -1.44 | -1.40 | -0.79 |
| Chad | -1.04 | -1.49 | -Inf | -Inf | -0.05 | -0.99 |
| Comoros | -1.06 | -Inf | -Inf | -Inf | -Inf | -0.72 |
| Dominican Republic | -1.09 | -1.47 | -Inf | -Inf | -1.50 | -0.85 |
| Gabon | -1.09 | -Inf | -Inf | -Inf | -Inf | -0.81 |
| Djibouti | -1.11 | -Inf | -0.05 | -Inf | -Inf | -1.05 |
| Kyrgyz Republic | -1.15 | -0.51 | -0.87 | -0.89 | -Inf | -1.40 |
| United Arab Emirates | -1.17 | -0.81 | -Inf | -0.86 | -1.26 | -1.20 |
| Azerbaijan | -1.21 | -0.79 | -0.87 | -Inf | -Inf | -1.17 |
| El Salvador | -1.24 | -1.19 | -Inf | -1.07 | -1.01 | -1.17 |
| Cayman Islands | -1.25 | -Inf | -Inf | -Inf | -Inf | -1.59 |
| Tajikistan | -1.29 | -0.23 | -0.84 | -Inf | -0.83 | -1.80 |
| Mauritania | -1.30 | -0.73 | -Inf | -Inf | -Inf | -1.33 |
| Curacao | -1.31 | -Inf | -Inf | -Inf | -Inf | -0.96 |
| Libya | -1.31 | -1.30 | -Inf | -Inf | -0.68 | -1.29 |
| Paraguay | -1.34 | -2.13 | -Inf | -Inf | -Inf | -1.07 |
| Kazakhstan | -1.35 | -0.67 | -1.44 | -Inf | -0.48 | -1.50 |
| Moldova | -1.36 | -Inf | -0.93 | -0.85 | -Inf | -1.39 |
| San Marino | -1.44 | -Inf | -Inf | -Inf | -Inf | -1.09 |
| Equatorial Guinea | -1.49 | -1.40 | -Inf | -Inf | -0.64 | -1.52 |
| Algeria | -1.54 | -2.42 | -1.39 | -Inf | -1.91 | -1.28 |
| Cote d'Ivoire | -1.66 | -Inf | -Inf | -Inf | -Inf | -1.27 |
| Belarus | -1.78 | -1.13 | -1.68 | -1.50 | -1.46 | -1.95 |
| Angola | -2.09 | -1.96 | -Inf | -Inf | -Inf | -1.89 |
| Uzbekistan | -2.45 | -1.63 | -Inf | -Inf | -1.63 | -2.45 |
| Turkmenistan | -3.06 | -Inf | -Inf | -Inf | -Inf | -3.07 |
| Sint Maarten (Dutch part) | -Inf | -Inf | -Inf | -Inf | -Inf | -Inf |
| Andorra |  |  |  |  |  |  |
| American Samoa |  |  |  |  |  |  |
| Cuba |  |  |  |  |  |  |
| Eritrea |  |  |  |  |  |  |
| Faroe Islands |  |  |  |  |  |  |
| Gibraltar |  |  |  |  |  |  |
| Greenland |  |  |  |  |  |  |
| Guam |  |  |  |  |  |  |
| Isle of Man |  |  |  |  |  |  |
| Liechtenstein |  |  |  |  |  |  |
| St. Martin (French part) |  |  |  |  |  |  |
| Monaco |  |  |  |  |  |  |
| Northern Mariana Islands |  |  |  |  |  |  |
| New Caledonia |  |  |  |  |  |  |
| Korea, Dem. People's Rep. |  |  |  |  |  |  |
| South Sudan |  |  |  |  |  |  |
| Syrian Arab Republic |  |  |  |  |  |  |
| Taiwan, China |  |  |  |  |  |  |
| Venezuela, RB |  |  |  |  |  |  |
| British Virgin Islands |  |  |  |  |  |  |
| Virgin Islands (U.S.) |  |  |  |  |  |  |
| Yemen, Rep. |  |  |  |  |  |  |
| French Polynesia |  |  |  |  |  |  |

## 8.3 Data Annex

Table. Relationship between Data Use in Academia and Data Sources from SPI.

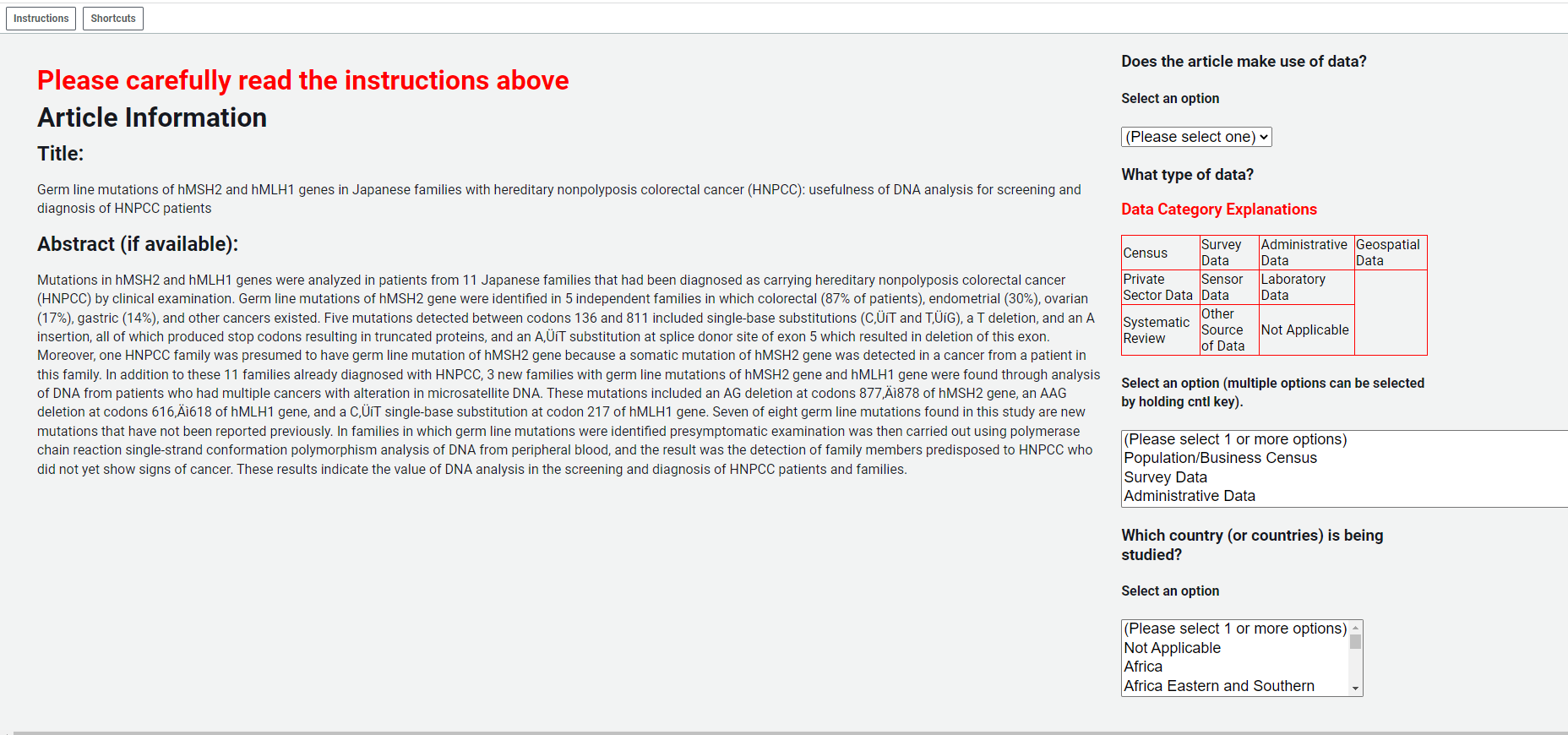
|  | (1) | (2) | (4) | (5) | (6) | (7) | (8) | (9) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Population Census in past 10 years | 0.02 |  |  |  |  |  |  |  |
|  | (0.07) |  |  |  |  |  |  |  |
| Agriculture Census in past 10 years |  | 0.002 |  |  |  |  |  |  |
|  |  | (0.04) |  |  |  |  |  |  |
| Household surveys: 2 or more in 10 years |  |  | 0.15\* |  |  |  |  |  |
|  |  |  | (0.07) |  |  |  |  |  |
| Agriculture Surveys: 2 or more in 10 years |  |  |  | -0.003 |  |  |  |  |
|  |  |  |  | (0.06) |  |  |  |  |
| Labor Force Surveys: 2 or more in 10 years |  |  |  |  | 0.08 |  |  |  |
|  |  |  |  |  | (0.07) |  |  |  |
| Health Surveys: 2 or more in 10 years |  |  |  |  |  | -0.09+ |  |  |
|  |  |  |  |  |  | (0.05) |  |  |
| Business/Establishments: 2 or more in 10 years |  |  |  |  |  |  | -0.06 |  |
|  |  |  |  |  |  |  | (0.06) |  |
| Complete Civil Registration and Vital Statistics System |  |  |  |  |  |  |  | 0.24+ |
|  |  |  |  |  |  |  |  | (0.13) |
| Log GDP per capita | 0.47\* | 0.48\* | 0.51\* | 0.48\* | 0.47\* | 0.46\* | 0.48\* | 0.49\* |
|  | (0.21) | (0.21) | (0.21) | (0.21) | (0.21) | (0.22) | (0.21) | (0.21) |
| Log Population | 0.34 | 0.35 | 0.32 | 0.35 | 0.39 | 0.32 | 0.36 | 0.33 |
|  | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.41) | (0.40) | (0.39) |
| N | 2009 | 2009 | 2009 | 2009 | 2009 | 2009 | 2009 | 2009 |
| Data from the World Bank's World Development Indicators (WDI) and SPI. WDI series codes include NY.GDP.PCAP.PP.KD, SP.POP.TOTL, IQ.SPI.OVRL,IQ.SPI.PIL1,IQ.SPI.PIL2, IQ.SPI.PIL3, IQ.SPI.PIL4, IQ.SPI.PIL5. Papers include all papers using data years 2000-2020.  \*\*\*=0.001 level  \*\*=0.01 level  \*=0.05 level  +=0.1 level | | | | | | | | |

## 8.4 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)