**Missing Evidence:**

**Tracking Academic Data Use around the World**

Brian Stacy, Lucas Kitzmüller, Xiaoyu Wang, Daniel Gerszon Mahler, and Umar Serajuddin[[1]](#footnote-2)

November 2023

Data-driven research on a country is key to producing evidence-based public policies. Yet little is known about where data-driven research is lacking and how it could be expanded. We propose a method for tracking academic data use by country of subject, applying Natural Language Processing (NLP) to open-access research papers. The model’s predictions produce country estimates of the number of articles using data that are highly correlated with a human-coded approach, with a correlation of 0.99. Analyzing over one million academic articles, we find that the number of articles on a country is strongly correlated with its GDP per capita, population, and the quality of its national statistical system. We identify data sources strongly associated with data-driven research and find that availability of subnational data appears particularly important. Finally, we classify countries into groups based on whether they could most benefit from increasing their data supply or the demand for its data. We find that the former applies to many low and lower-middle income countries while the latter applies to many upper-middle and high-income countries.

Keywords: Data, academia, research, natural language processing  
  
JEL codes: C45, C52, O30

# 1 Introduction

In recent decades, the amount of data produced has exploded, generating boundless opportunities for policies to improve people’s lives (World Bank 2021). Though data can be valuable in their raw form, the full value of data is only realized when they are analyzed to create insights, and these insights are converted to public policies or increased accountability.

Researchers have a vital role to play in this regard. Many researchers spend countless hours digesting data, using data to create new knowledge, and communicating this knowledge with the intent of impacting public discourse and public policies. There are numerous examples of data-driven analyses having real and important impacts on people’s lives (Jolliffe et al. 2023). One example from Brazil explicitly looks at researchers’ ability to influence policy outcomes. There, evidence from 2,150 municipalities found that informing municipal mayors of research findings on the effectiveness of a simple policy change, increased the probability that their municipality implemented the policy by 10 percentage points (Hjort et al. 2021).

Without research, there is a risk that the return of data to society will be reduced, and policies to improve lives unrealized. Yet very little is known about where there is missing data-driven evidence and how governments best can stimulate an evidence-base for local decision makers. This paper attempts to fill these gaps by addressing two questions: (1) Which countries are the subject of research papers using data? (2) How can countries increase its national evidence base? We focus on data-driven research due to the increasing importance of data for policy making and the specific policies that are needed to increase the supply and demand of data, such as boosting statistical capacity and improving data literacy.

To answer the first question, we introduce a new method for measuring data use in research articles based on one million English-language articles spanning 216 countries and various academic fields. These articles are made available by Semantic Scholar Open Research Corpus (s2orc), which has digitized millions of research papers worldwide and made its raw text accessible via APIs (Lo et al. 2020). With the aid of Amazon Mturk workers, we manually code 900 of these articles as using data or not, upon which we train a natural language model to predict the coding of the Mturk workers (Devlin et al. 2018). The model achieved an 87% out-of-sample accuracy rate, and when the articles are aggregated to the country level, the model had a correlation of 0.99 with the number of articles classified by Mturk workers. The model was then applied to 1 million academic articles from the s2orc database from 2000-2020. The model estimates the amount of data-driven research on a country regardless of where the researchers may be located, *not* the amount of data-driven research by the citizens of a country. We argue that the former is the relevant quantity to understand the evidence base available to national decision makers.

We find that data-driven research is strongly correlated with GDP per capita and population, which together account for around 75% of the variation across countries. High-income countries are the subject of nearly 50% of all papers using data despite representing only around 15% of the world's population, while low-income countries, comprising approximately 10% of the world population, account for only about 5% of articles using data.

To answer the second question – how countries can increase its national evidence base – we first establish that a country’s statistical capacity is predictive of data research even after controlling for population and GDP, and articles not using data (which we use as a proxy for the general research interest in the country). To understand which part of a country’s statistical capacity is most important for increasing data-driven research, we explore the data sources most related to academic data use. We find that the availability of geospatial data at the first administrative level is associated with a 1.1% increase in data use, that a population census in the past decade is associated with a 0.3% increase in data use, and that two or more labor force (agriculture) surveys over the past 10 years is associated with a 0.4% (0.2%) increase in data use. Though we are unable to establish these links causally, there are concrete data products that governments could supply to likely increase the evidence-base at its disposal.

Boosting the supply of data is one way for a country to increase the data-driven research it is subject to, another is to increase the demand for its data. This is particularly relevant for countries that have already invested in relevant data products but are nonetheless subject to relatively little data-driven research. These are cases where existing data are underutilized and where it may be relevant to make existing data more accessible to researchers and possibly boost data literacy in the country. To explore this distinction between boosting data supply and data demand, and building upon Porteous (2020), we classify countries into four groups: *Deserts* have little data demand and little data supply, *swamps* have high data supply but little data demand, *oases* have high data demand but little data supply, and *lakes* have high data demand and high data supply. Nearly two-thirds of low-income countries and countries in Sub-Saharan Africa are oases, suggesting that these countries are getting relatively large evidence from their data supply and that they have relatively little issue with lack of demand, but that they could benefit from increasing the data available to researchers. By contrast, nearly half of countries in Europe are data swamps, suggesting a priority on increasing the use of existing data.

Previous research have highlighted gaps between countries in economic research output and noted that richer countries are the subject of more economic research. 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. Cameron, Mishra, and Brown (2016) and Sabet and Brown (2018) extend this to note that impact evaluations are highly uneven across countries as well. Phillips and Greene (2022) show conflict research is biased towards Western countries, while Courtioux et al. (2022) show that academic research is highly related to public investments in scientific research.

We contribute to the literature by applying NLP to improve our understanding of what countries are under-researched. Using NLP allows us to go beyond the existing literature in three ways: (1) scale up the sample size and look at all fields of interest rather than only economics, (2) identify papers that use data, which is key to understanding whether data demand or data supply could be explaining a country’s lack of research, and (3) indicate measures countries can take to increase data research.

The remainder of the paper is structured as follows. Section 2 discusses our data sources, section 3 details our methodology, section 4 introduces a theoretical framework, section 5 presents our empirical results, section 6 conducts robustness checks, and section 7 concludes.

# 2 Data

Our data source of academic articles is the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al. 2020). The corpus contains more than 130 million English language academic papers across multiple disciplines. 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.

We placed some restrictions on the articles to make them usable and relevant for our purposes. 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 date 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 (i.e., suitable for digital processing) PDF, and around 26% of those 30 million are eliminated when removing articles without an abstract. Second, only articles from the year 2000 to 2020 were considered. This restriction eliminated an additional 9% of the remaining articles. Finally, articles from the following fields of study were excluded, as we aim to focus on fields that are likely to use data produced by countries’ 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 the remaining articles. From an initial corpus of 136 million articles, this resulted in a final corpus of around 10 million articles.

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 1. 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 | 0.56 | 0.64 | 0.30 | 28,571 | 2.8 |
| Economics | 0.79 | 0.68 | 0.28 | 62,241 | 6.0 |
| Medicine | 0.96 | 0.85 | 0.10 | 840,920 | 81.0 |
| Political Science | 0.42 | 0.33 | 0.34 | 26,185 | 2.5 |
| Psychology | 0.75 | 0.70 | 0.14 | 44,191 | 4.3 |
| Sociology | 0.90 | 0.33 | 0.25 | 35,640 | 3.4 |

# 3 Empirical Strategy

The empirical approach employed in this project utilizes text mining with Natural Language Processing (NLP). The goal of NLP is to extract structured information from raw, unstructured text. In this project, NLP is used to extract the country of study and whether the paper makes use of data. We will discuss each of these in turn.

To determine the country or countries of study in each academic article, two approaches are employed based on information found in the title, abstract, or topic fields. The first approach uses regular expression searches based on the presence of ISO3166 country names. A defined set of country names is compiled, and the presence of these names is checked in the relevant fields. This approach is transparent, widely used in social science research, and easily extended to other languages. However, there is a potential for exclusion errors if a country’s name is spelled non-standardly.

The second approach is based on Named Entity Recognition (NER), which uses machine learning to identify objects from text, utilizing the spaCy Python library. The Named Entity Recognition algorithm splits text into named entities, and NER is used in this project to identify countries of study in the academic articles. SpaCy supports multiple languages and has been trained on multiple spellings of countries, overcoming some of the limitations of the regular expression approach. If a country is identified by either the regular expression search or NER, it is linked to the article. Note that one article can be linked to more than one country.

The second task is to classify whether the paper uses data. A supervised machine learning approach is employed, where 3500 publications were first randomly selected and manually labeled by human raters using the Mechanical Turk service (Paszke et al. 2019).[[2]](#footnote-3) To make sure the human raters had a similar and appropriate definition of data in mind, they were given the following instructions before seeing their first paper:

*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 any 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 click multiple options.[[3]](#footnote-4)*

*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, select all countries that are discussed in the paper*.

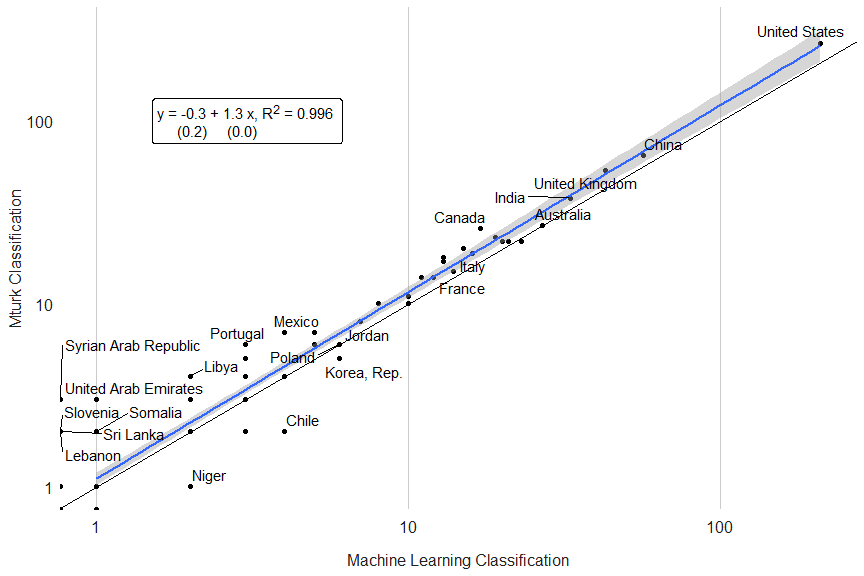
*We expect between 10 and 35 percent of all articles to use data.*

An image of the screen facing the MTurk workers when classifying an article is presented in Figure A.1 in the Appendix. The median amount of time that a worker spent on an article, measured as the time between when the article was accepted to be classified by the worker and when the classification was submitted was 25.4 minutes. If human raters were exclusively used rather than machine learning tools, then the corpus of 1,037,748 articles examined in this study would take around 50 years of human work time to review at a cost of $3,113,244, which assumes a cost of $3 per article as was paid to MTurk workers.

A model is next trained on the 3,500 labelled articles. We use a distilled version of the BERT (bidirectional Encoder Representations for transformers) model to encode raw text into a numeric format suitable for predictions (Devlin et al. (2018)). BERT is pre-trained on a large corpus comprising the Toronto Book Corpus and Wikipedia. The distilled version (DistilBERT) is a compressed model that is 60% the size of BERT and retains 97% of the language understanding capabilities and is 60% faster (Sanh, Debut, Chaumond, Wolf 2019). We use PyTorch to produce a model to classify articles based on the labeled data. Of the 3,500 articles that were hand coded by the MTurk workers, 900 are fed to the machine learning model. 900 articles were selected because of computational limitations in training the NLP model. A classification of “uses data” was assigned if the model predicted an article used data with at least 90% confidence.

The performance of the models classifying articles to countries and as using data or not can be compared to the classification by the human raters. We consider the human raters as giving us the ground truth. This may underestimate the model performance if the workers at times got the allocation wrong in a way that would not apply to the model. For instance, a human rater could mistake the Republic of Korea for the Democratic People’s Republic of Korea. If both humans and the model perform the same kind of errors, then the performance reported here will be overestimated.

The model was able to predict whether an article made use of data with 87% accuracy evaluated on the set of articles held out of the model training. The correlation between the number of articles written about each country using data estimated under the two approaches is given in the figure below. The number of articles represents an aggregate total of research output using data for each country summed from the corpus of papers that were not used to train the model. The Pearson correlation between the human raters and the NLP predictions is 0.996.

*Figure 1. Comparison of Human Classifications of Data Use to NLP Predictions *

*Note:* The horizontal axis shows the number of articles on a country using data as predicted by the natural language processing model. The vertical axis shows the number of articles on a country using data as classified by the human raters (Mturk workers).

To make the performance of the model more concrete, we consider the output returned for three example articles. These articles were not in the training set of articles, so the model has not previously seen them.

The first article, Perlman (2009), titled, “The Legal Ethics of Metadata Mining” is a law essay examining the ethics of metadata mining. This article has data (or metadata) as the subject of the article, but does not actually use data to perform analysis, making it potentially difficult to machine classify. In the authors’ judgement, this article should not be classified as using data. The NLP model also evaluated that this article did not use data.

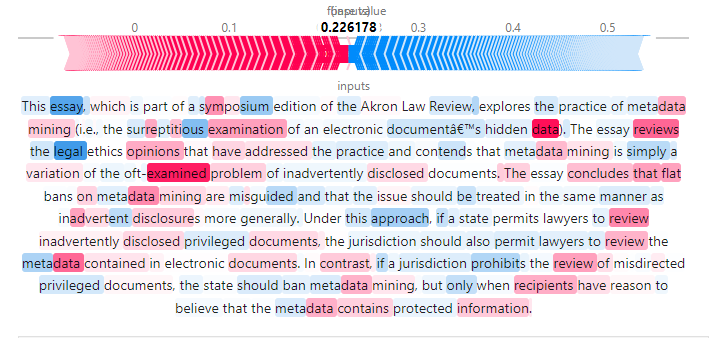
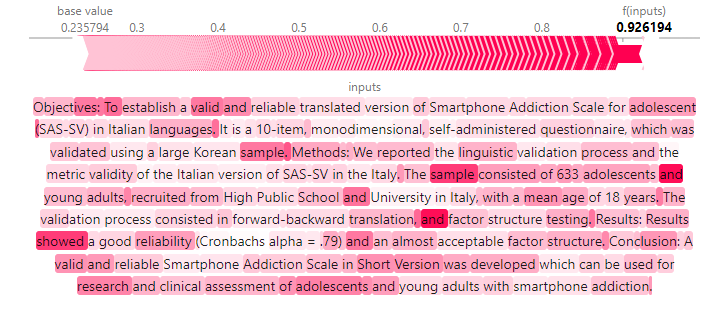
Figure 2 shows words that the model viewed as indicating data use (in red) and likely indicating an article does not use data (in blue) using the SHAp (SHapley Additive exPlanations) package in python (Lundberg and Lee 2017). While the model picked up keywords like data and examined (highlighted in red), indicating data use, the NLP algorithm also picked up other key words such as legal and review, which had a reduced likelihood of using data. On balance the predictions of the NLP model indicated the article did not use data. Using an alternative approach, such as flagging key words in the article abstract like data, would have flagged this article as using data and given the incorrect classification.

The second article, De Pasquale, Concetta, Federica Sciacca, and Zira Hichy (2017), titled, “Italian validation of smartphone addiction scale short version for adolescents and young adults (SAS-SV)”, was classified as using data. In the authors’ judgement, this was a correct classification. The article picked up on key words such as sample, showed, scale, and numeric values, indicating data use (highlighted in red).

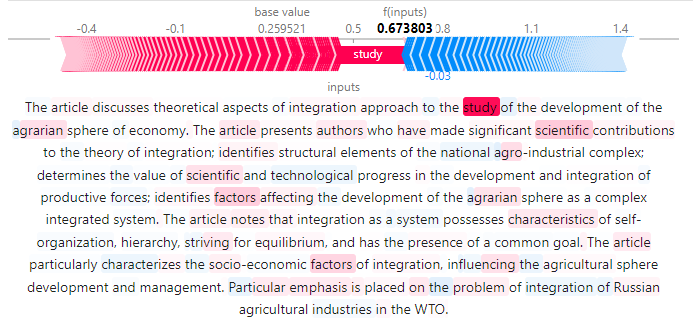
The third article, Smirnov, Anatoly A., and Irina V. Stukova (2015), titled, "Determinants of integration approach in the agrarian sphere development in contexts of transformation." was given a probability of using data of 0.67, which was below the threshold of 0.9 used to assign a label of data use. However, on the model thought the article more likely than not to use data. In fact, the article does not use data, indicating that model can be ambiguous about the status of an article.

Figure 2. Article Examples

1. No Data Use (b) Data Use

*(c) Model Unsure*



*Note*: The left panel is a sample from Perlman, Andrew M. “The Legal Ethics of Metadata Mining.” Akron Law Review 43 (2009). The right panel is a sample from De Pasquale, Concetta, Federica Sciacca, and Zira Hichy. “Italian validation of smartphone addiction scale short version for adolescents and young adults (SAS-SV).” Psychology 8, no. 10 (2017). The bottom panel is a sample from Smirnov, Anatoly A., and Irina V. Stukova. "Determinants of integration approach in the agrarian sphere development in contexts of transformation." *Review of European studies* 7, no. 8 (2015): 8.

After applying the natural language processing model to around 1 million articles from the s2orc corpus, we compare our estimated number of articles produced using data to previous estimates in the literature. Das et al. (2013) use a corpus of more than 76,000 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, the correlation between country output is still 0.62 (Figure A.2a).

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 (Figure A.2b). A third comparison can be made to the number of scientific and technical journal articles produced by the National Science Foundation (NSF) (National Science Board, National Science Foundation (2019)). The NSF counts the scientific and engineering articles published in 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). Comparing the NSF data with estimates from our NLP model for 2018 gives a correlation of around 0.9 (Figure A.2c).

# 4 Theory

Building on a model introduced in Porteous (2022), we motivate analysis between the total amount of academic research using data and other development outcomes as follows. In a simple model with competitive members of academia, academics produce research using data, where benefits to that research scale with the size of the population of the country. That is, research that is about countries with greater populations produce higher benefits. This can be motivated if researchers choose to produce output that affects as many individuals as possible or else the prestige of academic work grows with the size of the country. As Porteous (2022) discusses, academics may also be interested in studying larger and more complex economies as the topics available to study (sectors and the policy environment) scales with the income level of a country. Academic research shows diminishing returns, where the payoffs decline depending on the quantity of existing research.

The marginal cost of additional research using data depends on a variety of factors, such as the stock of existing data available for research in the country and other factors, such as differences in the cost of field work across countries. Researchers then compete so that the net benefit of additional research is identical across countries, so that for two countries :

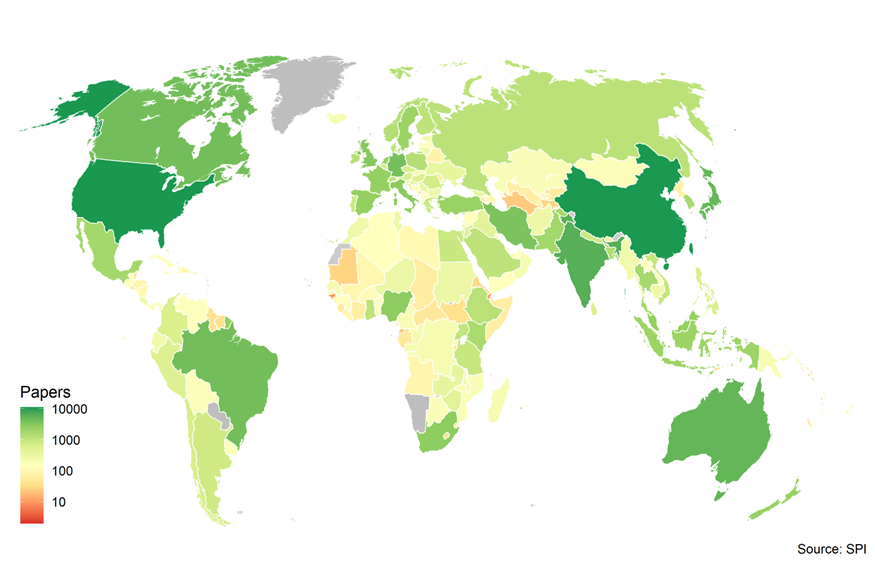
Where B(.) is a benefits function, C(.) represents costs, is the population size, is a measure of economic size, is the stock of economic research, and represents other factors affecting the cost of producing the research. Equation (1) motivates a model where the total amount of research in a country follows a Cobb-Douglas production function that depends on these factors listed above, where:

Assuming that , Equation (2) implies that there is diminishing marginal returns to the size of the country and the size of the economy. Additionally, as the factors affecting cost increase, the model predicts total research output using data to decrease. The model in log form is:

Equation (3) can be estimated using OLS, and the coefficients can be interpreted as the elasticities of population size, economic output, and factors affecting costs of gathering data with respect to academic output using data. As a measure of the cost of gathering data for research, we use as a proxy variable the World Bank’s Statistical Performance Indicators (SPI), a measure of the performance of statistical systems for 174 countries (Dang et al. 2023, Cameron et al. 2021). Scores on the SPI are on a scale of 0-100, where countries scoring near 100 are the best performing systems and countries scoring closer to zero are the lowest performing systems. Because the term for in equation (3) is structured as the elasticity with respect to costs of data collection, and because SPI scores are formulated so that better performing systems score higher, the coefficient on the SPI score in a regression will have the opposite sign as .

# 5 Results

Using the NLP model, the number of articles using data produced for each country is shown in Figure 3. 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). India, Australia, and Japan are third, fourth, and fifth with 6,481, 5,463, and 5,300 papers respectively.

*Figure 3. Number of Articles using Data by Country of Subject (2000-2020)* 

Five countries are the subject of more than 25% of all academic output using data. However, these countries also 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 are subject of less than 1% of output. These shares match roughly to the population size of each of these groups.

*Figure 4. Number of Articles Using Data by Income Group and Region*

|  |  |
| --- | --- |
| *a) By income group* | *b) By region* |
|  |  |
| *Note*: Panel (a) shows the total number of articles classified as using data by income group. Panel (b) shows the number of articles classified as using data by region per million people. | |

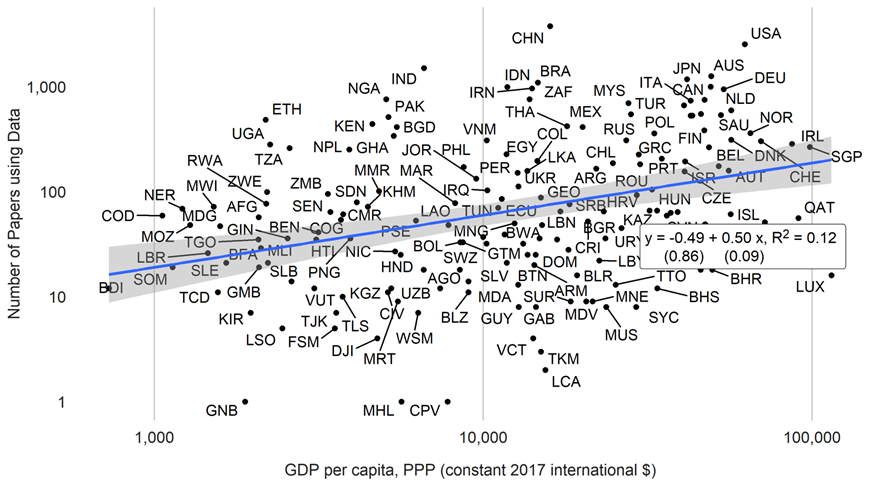
There is much more imbalance when looking at income groups (Figure 4a). High income countries are the subject of nearly 50% of all papers using data from 2000-2020, despite only making up around 17% of the world’s population. Despite making up around 1/10th of the world population, low-income countries only account for around 5% of articles using data. When looking by region (Figure 4b), Europe and Central Asia had the largest number of articles using data on a per capita basis with around 430 quantitative articles per million. North America is second with around 410 per million. South Asia is the subject of the fewest number of papers per million with only around 59 articles produced per million persons, less than half than Sub-Saharan Africa, the region with the second lowest quantitative articles per million. Table A.1 in the Appendix contains the total number of articles using data as well as the total number per million at the country level.

## 5.1 Relationships to Development Outcomes

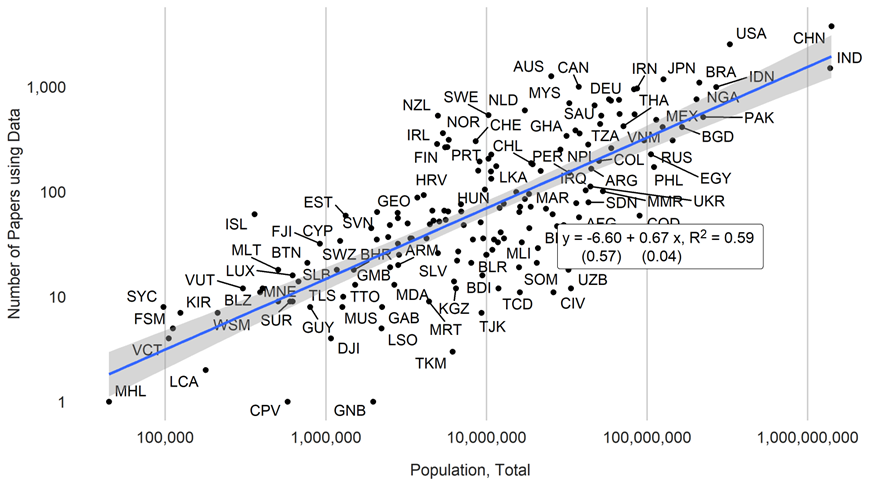
We estimate equation (3) using a dataset containing 167 countries with data available on all key variables. The number of articles is the three-year average count of all articles produced using data for each country between 2017 and 2019. A three-year average is used rather than a single year value for 2019, because the production of articles can be volatile for a single year observation. Population, GDP per capita, PPP, and SPI scores are all for the year 2019. There is a strong relationship between the number of articles produced using data and GDP per capita, population, and the overall SPI score of a country. Bivariate regressions of log number of papers using data on log GDP, log population, and the SPI overall score are shown in Figure 5.

*Figure 5. Relationship between Papers Using Data and Development Outcomes*

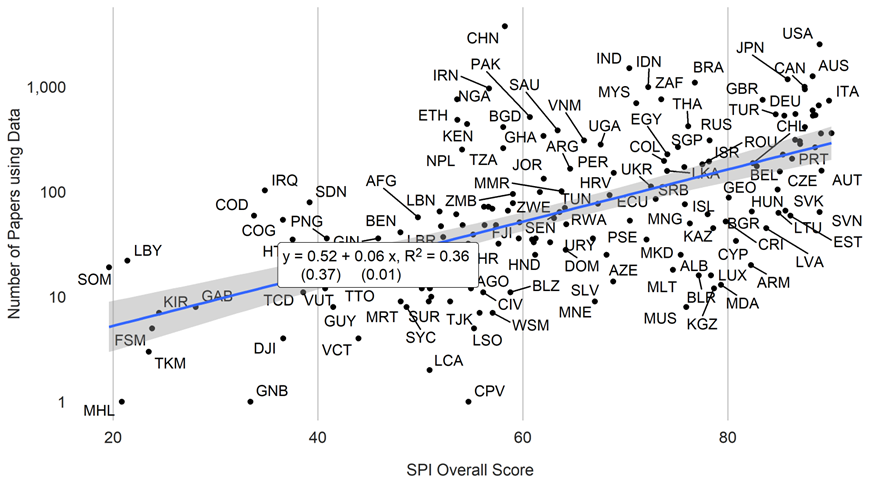
1. *GDP per capita*



1. *Population*



*(c) SPI Overall Score*



*Note*: Data on GDP, population and SPI are retrieved from the World Bank’s World Development Indicators

For GDP per capita, the figure indicates an elasticity of 0.5, meaning that a 10% increase in GDP is associated with a 5% increase in academic articles using data. The elasticity with respect to population is 0.67, and population alone explains 59% of the variation in academic articles using data. The overall SPI score is also 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 the65th percentile, translates into around a 0.6% increase in the number of articles using data.

Table 2 shows coefficients from a regression of the log number of academic papers using data on log GDP per capita, log population, and the SPI scores. The first column shows a specification including just log GDP per capita and log population. These two indicators alone explain roughly 75% of the variation in the production of articles using data. The second column includes the SPI overall score as an additional control. The performance of a country’s statistical system 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.

It is possible that omitted variables are behind this result. For example, it could be the case that countries that are of general interest to the research community also happen to have high statistical performance. Das et al. (2013) and Porteous (2022) control for a range of variables which may influence the relationship sought after, such as whether English is a first language, international tourist arrivals, etc. We leverage that we have an estimate of the number of articles not using data – referred to as qualitative articles – to control for all factors that are likely to influence overall research output in a country unrelated to statistical performance. The number of qualitative articles is calculated as the number of total articles minus any articles using data, based on our NLP model.

Countries that tend to be over-represented in quantitative articles also tend to be over-represented in qualitative articles, conditional on GDP per capita and population (Figure A.3). The elasticity is around 0.6, suggesting that a 1% increase in the number of qualitative articles associated with a 0.6% increase in the number of articles using data. We take this to imply that quantitative articles do not crowd out qualitative articles and vice versa. Returning to Equation (2), one prediction from the model is that countries with particularly high costs related to using data, (such as a lower statistical performance score), should see reductions in the amount of quantitative research, but it would not follow that it would see lower levels of qualitative research. This would predict the share of quantitative articles to total articles increasing in SPI score. Contrary to this, there is almost no relationship between SPI overall scores and the share of qualitative papers to total papers produced overall (Figure A.4). This suggests to us that some omitted variable may in part be driving the impact of statistical performance on academic articles using data, and that the number of qualitative articles can control for such possibly omitted factors.

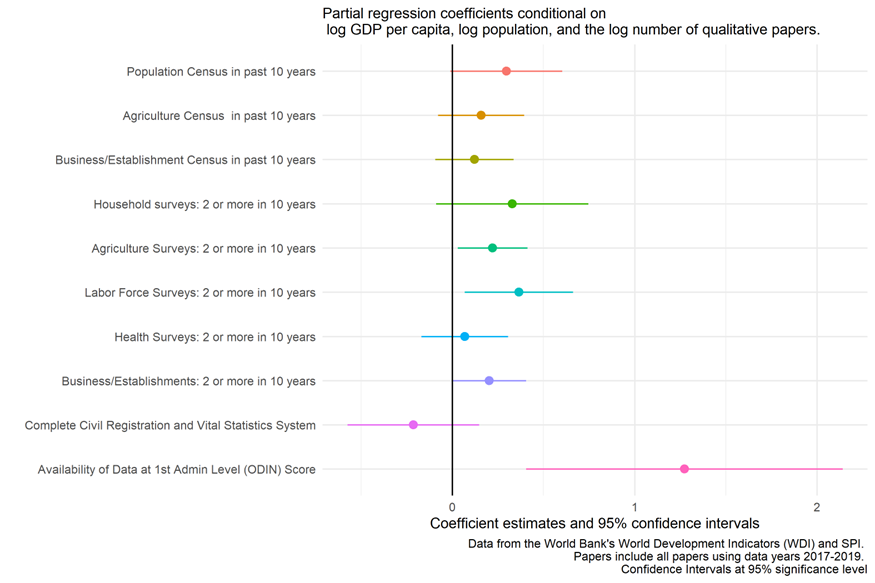
The third column of Table 2 additionally controls for the log number of qualitative articles. The SPI overall scores are still statistically significant at the 5% level, as is country population. Column 4 examines the relationship with one of the pillars of the SPI, data sources, which covers the availability of recent censuses, surveys, academic data, and geospatial data. The data sources index from the SPI is also correlated with data use in academia, conditional on log GDP, population, and the number of qualitative papers. Similar regressions using the other pillars of the SPI, such as data services and data infrastructure are not associated with the number of articles using data in a similar regression, suggesting that it is the availability of particular data sources that is helpful in boosting academic data use.

*Table 2. Relationships between Number of Papers Using Data and Statistical Performance*

|  | (1) | (2) | (3) | (4) |
| --- | --- | --- | --- | --- |
| (Intercept) | -12.41\*\*\* | -10.66\*\*\* | -3.96\*\*\* | -3.57\*\* |
|  | (0.82) | (0.95) | (1.14) | (1.17) |
| Log GDP per capita | 0.58\*\*\* | 0.36\*\*\* | 0.12+ | 0.10 |
|  | (0.05) | (0.07) | (0.07) | (0.08) |
| Log Population | 0.70\*\*\* | 0.63\*\*\* | 0.29\*\*\* | 0.29\*\*\* |
|  | (0.03) | (0.04) | (0.05) | (0.05) |
| SPI Overall Score |  | 0.02\*\*\* | 0.01\* |  |
|  |  | (0.01) | (0.01) |  |
| Log Qualitative Papers |  |  | 0.64\*\*\* | 0.65\*\*\* |
|  |  |  | (0.08) | (0.08) |
| SPI Data Sources Score |  |  |  | 0.01\* |
|  |  |  |  | (0.00) |
| Observations | 168 | 168 | 168 | 168 |
| R2 | 0.751 | 0.774 | 0.852 | 0.852 |
| R2 Adj. | 0.748 | 0.770 | 0.849 | 0.849 |
| AIC | 414.8 | 400.1 | 331.1 | 330.7 |
| BIC | 424.2 | 412.6 | 346.8 | 346.3 |
| RMSE | 0.82 | 0.78 | 0.63 | 0.63 |
| *Note:* GDP, population and SPI data are retrieved from the World Bank's World Development Indicators. Regressions include all papers using data from 2017-2019. Standard errors are robust to heteroskedasticity. \*\*\*=0.001 level, \*\*=0.01 level, \*=0.05 level, +=0.1 level. | | | | |

Using more detailed data from the SPI makes it possible to assess how the availability of specific data sources relates to academic output. Figure 6 shows the linear regression coefficient for the availability of each of the ten data sources considered by the SPI, conditional on log GDP per capita, log population, and log of qualitative papers. 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.

*Figure 6. Relationship between Number of Papers Using Data and Data Sources*



*Note*: Results from cross-sectional regression controlling for log population, log GDP per capita, and log number of qualitative papers. GDP and population data are retrieved from the World Bank’s Development Indicators, while the data source variables are from the Statistical Performance Indicators. The sample includes all papers using data between 2017 and 2019. Confidence intervals are at the 95% significance level. Full regression results are presented in Table A.2.

The regression estimates indicate that the most important source is data availability at the Admin 1 level, which is associated with a 1.3% increase in academic output using data. The availability of a population census in the past 10 years is associated with a .3% increase in papers using data, which is significant at the 10% level. The availability of two or more labor force (agriculture) surveys over the past 10 years is associated with a 0.4% (0.2%) increase in data use, both significant at the 5% level. Though the magnitude of these effects may appear small, they likely do not reflect the full potential that investing in data sources can have, as some of the countries that have relevant data sources are faced with low use of its products for a variety of reasons. To explore this further, it is instructive to consider both the supply and demand for data.

## 5.2 Data Supply vs. Data Demand: A Story of Data Deserts, Oases, Swamps, and Lakes

The prior analysis suggested that countries can increase the number of data-driven academic articles it is subject to by increasing its supply of relevant data sources. Yet some countries may already be performing quite well in terms of the data supply it is producing. For such countries, the primary obstacle preventing more data research may not be the supply of data but the use of existing data. There are many reasons why existing data may not be used, including data being inaccessible to researchers and data being difficult to understand and process, for example due to lack of metadata (World Bank 2021). Other reasons relate lack of data literacy among researchers, low trust in data from the national statistical system, and lack of infrastructure to access and use the data, which may be particularly an issue of low- and middle-income countries.

The policies needed to boost data supply differ from those needed to boost data demand. Boosting the data supply of the national statistical system requires more financing for national statistical offices, better technical capacity of its staff, and at times better statistical laws to ensure their independence from other government bodies. Policies to boost data demand, on the other hand, involve financially cheap (but not necessarily politically cheap) wins, such as making data more accessible for researchers. It also involves increasing the data literacy of academics. Given the non-overlapping policy responses, it is useful for countries to know whether data demand or data supply is their weakest link in the data value chain. Below, we classify countries by whether they are primarily constrained by data supply, data demand, neither or both.

Building upon Porteous (2020), we classify countries into four groups: *Data deserts* have little data demand and little data supply, *data swamps* have high data supply but little data demand, *data oases* have high data demand but little data supply, and *data lakes* have high data demand and high data supply. For assessing data supply, we again turn to the availability of various data sources, such as recent population and business censuses, household and health surveys, Civil Registration and Vital Statistics (CRVS), and sub-national geospatial data (for details see Dang et al., 2023). We measure data use as the residual from the regression of number of articles on the country using data on log population, log GDP per capita, and the average number of qualitative papers. The control for the number of qualitative papers once again serves as a proxy for general research interest and research environment unrelated to data.

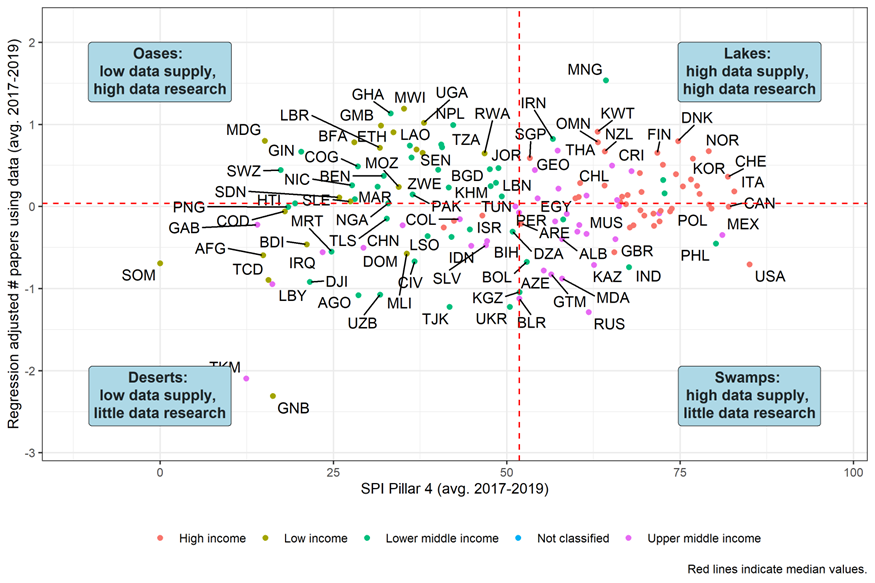
We classify countries by whether they are below/above the median of our indicators of data supply and data use (Figure 9). Since we use median values in the two dimensions to create the categories, there are about as many countries belonging to each group. Yet there are large discrepancies by region and income group (Table 3). The poorest countries and poorest regions are more likely to be data deserts or data oases. These categories apply to more than 80% of low- and lower-middle income countries, 93% of countries in Sub-Saharan Africa, and only 7% of high-income countries. Thus, poorer countries and poorer countries tend to have lower data supply.

Low-income countries and countries in Sub-Saharan Africa are twice as likely to be oases than deserts. In fact, when looking at the map, it is evident that a large part of the world’s data lakes, are to be found in Sub-Saharan Africa. This suggests that, despite data scarcity, a significant volume of quantitative articles is already produced on these countries, when factoring in their population size, income, and attention in non-quantitative scholarship. These countries may benefit most from bolstering their data supply. This applies for example to Uganda, Ghana, Nepal, and Malawi. By contrast, nearly all upper-middle income countries and countries in Latin America who have low data supply are deserts rather than oasis. That is, they face both data supply and data use constrains. This concerns countries such as Libya, Turkmenistan, and El Salvador.

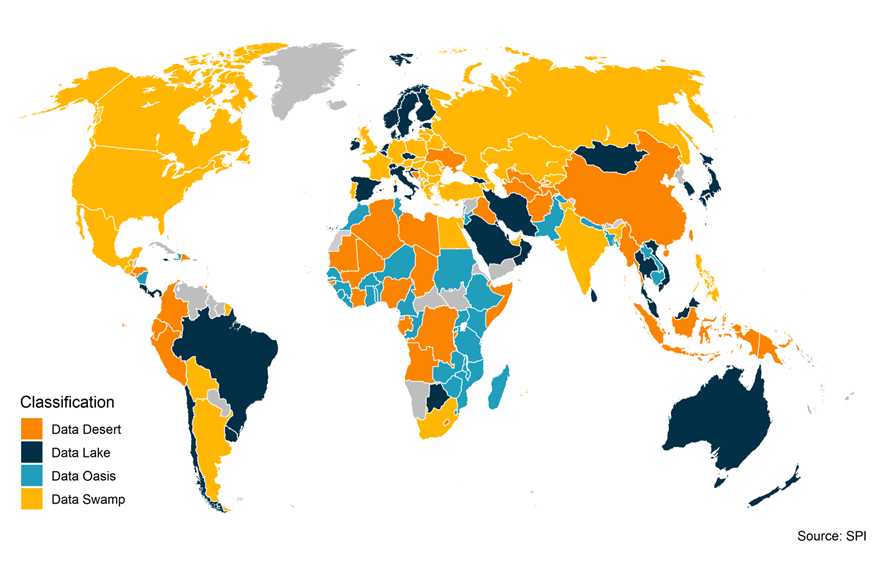
High-income countries and countries in East Asia & the Pacific are more likely to be a lake than a swamp. By contrast, upper-middle income countries and countries in Europe and North America are more likely to be a swamp than a lake. Hence, the latter countries may have successfully improved data supply but are lacking in data use. This concerns countries such as Russia, the UK, France, Germany, and the U.S. The Scandinavian countries are all classified as data lakes, meaning they have high supply and high use of their data.

Figure 7*. Relationship between Data Use and Data Supply*

1. *Country Scatterplot*



1. *Map*



Note: Panel (a) plots on the vertical axis the residuals from a regression of the average number of data papers between 2017 and 2019 on log GDP, log population, and the log number of papers not using data (qualitative papers). The horizontal axis plots SPI Pillar 4, which tracks the availability of recent censuses, surveys, Civil Registration and Vital Statistics, and subnational data (for details see Dang et al., 2023). Red dashed lines represent median values.

*Table 3. Data Deserts, Oases, Lakes, and Swamps by Region and Income Group*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Data Deserts | Data Oases | Data Swamps | Data Lakes | Number of countries |
| Region |  |  |  |  |  |
| East Asia & Pacific | 29% | 12% | 6% | 53% | 17 |
| Europe & Central Asia | 11% | 0% | 58% | 31% | 45 |
| Latin America & Caribbean | 37% | 11% | 21% | 32% | 19 |
| Middle East & North Africa | 33% | 22% | 11% | 33% | 18 |
| North America | 0% | 0% | 100% | 0% | 2 |
| South Asia | 17% | 50% | 17% | 17% | 6 |
| Sub-Saharan Africa | 31% | 62% | 5% | 3% | 39 |
| Income group |  |  |  |  |  |
| Low-income | 35% | 65% | 0% | 0% | 20 |
| Lower-middle income | 31% | 49% | 11% | 9% | 45 |
| Upper-middle income | 33% | 0% | 44% | 22% | 36 |
| High-income | 7% | 0% | 38% | 56% | 45 |
| Total | 25% | 24% | 26% | 25% | 146 |

*Note*: Share of countries within a region or income group that are classified as data deserts, oases, swamps, or lakes. For example, 29% of the 17 countries in East Asia & Pacific are classified as data deserts.

In Figure A.5 we show the classification when using number of quantitative articles per capita (unadjusted for GDP or number of qualitative articles) as the measure of data use. With this classification, high-income, often OECD, countries are doing well in generating articles using data. Many of the *data deserts* are in low-income or conflict affected countries. While the per capita measure is useful for an overview of how countries are performing in producing articles using data, it is not always useful as a policy guide given that it is partially driven by GDP per capita and research interests, which are often outside the control of policy makers in the statistical system.

# 6 Robustness Checks

In this section we offer some robustness checks to our main results. First, we estimate our main regressions using country fixed effects and year dummy variables rather than controlling for qualitative articles. Second, we investigate our results if we exclude medical papers.

## 6.1 Panel analysis

Our analysis may be subject to some omitted variable bias if countries with high statistical capacity share commonalities unrelated to their number qualitative articles. Below, rather than controlling for the number of qualitative articles, we add country fixed effects and year dummy variables to our OLS regressions, so that only within-country variation over time is used to identify our main relationships. Standard errors are clustered at the country level. Note that this specification may remove relevant variation if some countries have a constant high output of papers using data due to a constant high statistical capacity and vice versa.

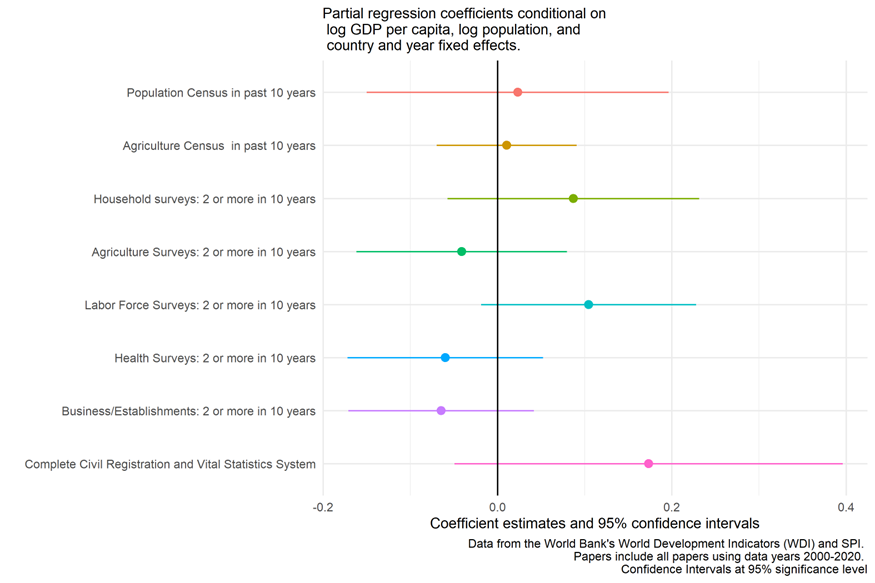
Using this specification, the relationship between academic output using data and log GDP per capita is similar as previously estimated, with an elasticity between around 0.4 and 0.6 depending on the specification (Table 4). The relationship with population is not statistically significant under this specification in most cases. The SPI overall scores are not statistically significant either under this specification, but because the SPI overall scores only are available starting in 2016, the time series is relatively short. To overcome this shorter time series, an “extended SPI” index is created that incorporates data from the older Statistical Capacity Indicator (SCI), which the World Bank has produced since around 2004.[[4]](#footnote-5) Using the longer time series in the extended SPI, a 10 unit increase in the extended SPI overall score is associated with an increase in academic output using data of around 0.1%. The estimated coefficient is statistically significant at the 5% level and similar in magnitude to the cross-sectional results that control for the number of qualitative articles. The extended SPI data sources score is also statistically significant at the 5% level.

*Table 4. Longitudinal Relationships between Number of Papers Using Data and Statistical Performance*

|  | (1) | (2) | (3) | (4) | (5) | (6) |
| --- | --- | --- | --- | --- | --- | --- |
| Log GDP per capita | 0.58\*\* |  |  | 0.51 | 0.58\*\* | 0.62\*\* |
|  | (0.20) |  |  | (0.33) | (0.22) | (0.22) |
| Log Population |  | 0.18 |  | 1.09 | 0.40 | 0.38 |
|  |  | (0.35) |  | (0.69) | (0.37) | (0.37) |
| SPI Overall Score |  |  | 0.00 | 0.00 |  |  |
|  |  |  | (0.00) | (0.00) |  |  |
| SPI Overall Score  (Extended Series) |  |  |  |  | 0.01+ |  |
|  |  |  |  |  | (0.00) |  |
| SPI Data Sources Score  (Extended Series) |  |  |  |  |  | 0.00+ |
|  |  |  |  |  |  | (0.00) |
| Observations | 1974 | 1974 | 661 | 661 | 1974 | 1974 |
| R2 | 0.957 | 0.956 | 0.991 | 0.991 | 0.957 | 0.958 |
| R2 Adj. | 0.953 | 0.952 | 0.988 | 0.988 | 0.954 | 0.954 |
| R2 Within | 0.025 | 0.001 | 0.001 | 0.013 | 0.036 | 0.039 |
| R2 Within Adj. | 0.025 | 0.000 | -0.001 | 0.007 | 0.035 | 0.037 |
| AIC | 1571.9 | 1621.2 | -280.9 | -285.1 | 1554.1 | 1548.5 |
| BIC | 2421.3 | 2470.6 | 492.0 | 496.8 | 2414.6 | 2409.0 |
| RMSE | 0.33 | 0.34 | 0.15 | 0.15 | 0.33 | 0.33 |
| *Note:* Data from the World Bank’s World Development Indicators (WDI) and SPI. 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. All regressions include standard errors clustered at the country level and country and year fixed effects. \*\*\*=0.001 level, \*\*=0.01 level, \*=0.05 level,+=0.1 level. | | | | | | |

Using a similar methodology as before, which combines SPI data with indicators available from the older SCI to create an extended data series back to 2004, partial regression coefficients are shown for the panel regression in the figure below. Using the panel regressions with country and year fixed effects, the coefficient for labor force surveys remains statistically significant at the 10% level. The coefficient however shrinks from around 0.35 in the cross-sectional regression to close to 0.15.

*Figure 8. Panel Relationships Between Number of Papers Using Data and Data Sources*



*Note*: Results from panel regression controlling for year and country-fixed effects, log population and log GDP. Data from the World Bank’s Development Indicators (WDI) and the Statistical Performance Indicators. The sample includes all papers using data between 2000 and 2020. Confidence intervals are at the 95% significance level.

## 6.2 Excluding medical articles

As noted previously, medical articles make up around 81% of all articles in our corpus. A concern could be that medical articles receive too much weight in our calculations of country scores. As a check, we compare our main results when dropping all medical articles. Except for countries with very few articles classified as using data, there is a very high correlation between the measures based on all fields and measures excluding medicine (Figure 9). This makes us comfortable that the high share of medical articles is not driving our results.

*Figure 9. Comparison Between All Papers Using Data and Non-Medical Papers Using Data*

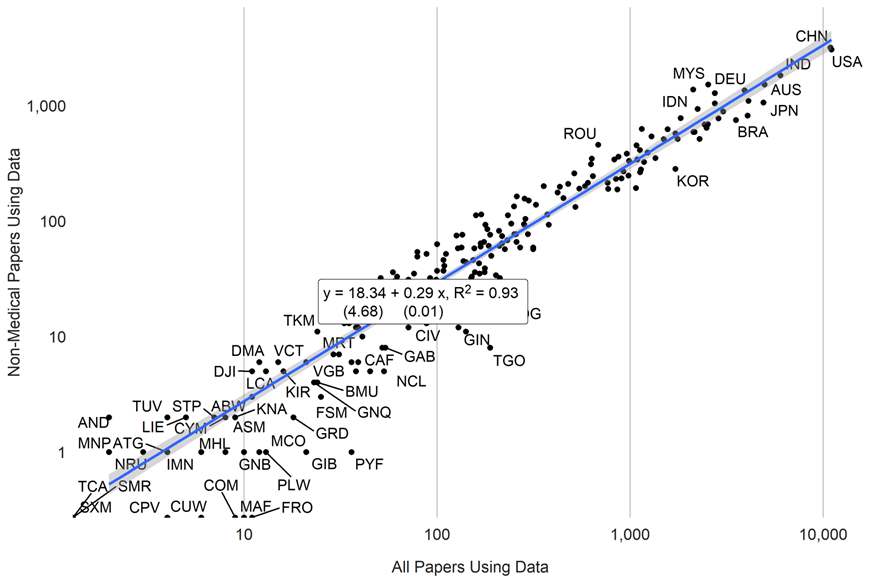
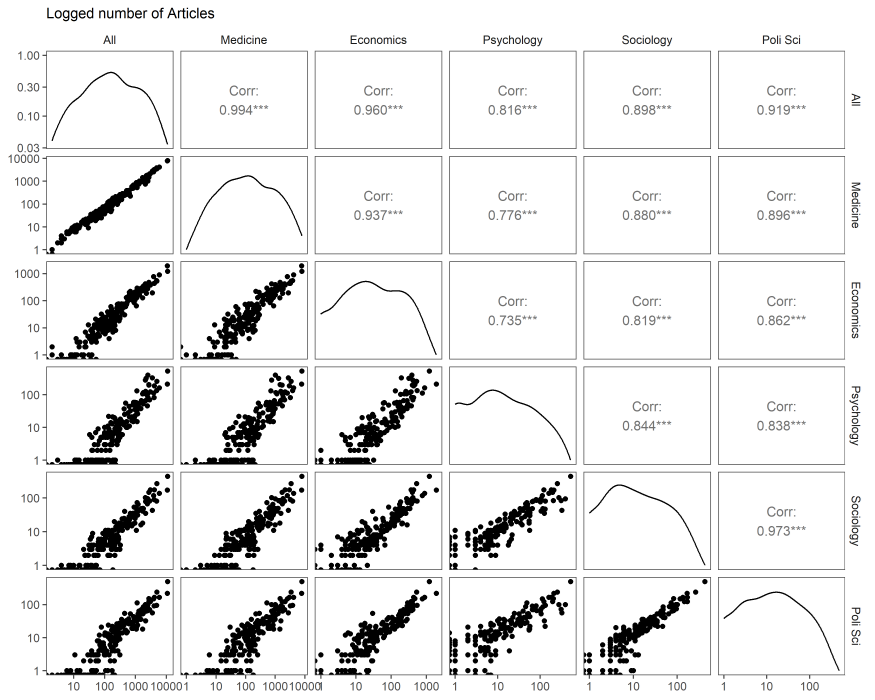


Figure 10 shows the relationships between the number of articles per country using data for all subjects and for each subject respectively between 2000 and 2020. In all cases, academic output using data is strongly correlated across subjects. The correlation between the number of medical articles using 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 0.97. Economics and Psychology have the lowest correlation (0.74).

*Figure 10. Correlation in Papers Using Data Across Subjects. 2000-2020.* 

# 7 Conclusion

This paper suggests a method for measuring data use by researchers by classifying academic articles using natural language processing (NLP). Our NLP model can classify whether articles use data with 87% accuracy and if aggregated to the country-level, match human rated country-level classifications of academic papers using data with a correlation of around 0.99. The obtained classification correlates strongly with alternative measures in the literature.

We apply the NLP model to around 1 million academic articles. We find that high-income countries are subject of a disproportionate share of academic output using data compared to their population. High-income countries account for around 50% of all articles, while making up only 17% of the world population. Low-income countries are the subject of only 5% of academic output using data, while making up 10% of the global population. GDP per capita and population are strong predictors of academic output, explaining around 75% of the variation between countries.

We also find that statistical performance – even after conditioning on population, GDP per capita – is a strong predictor of academic output using data, particularly in the case of data sources available. We explore the data sources that are associated with data use, and find population censuses, geospatial data, agricultural surveys, and labor force surveys to be most relevant. Finally, we classify countries into four clusters, indicating whether they could most benefit from extending their availability of core data products or make existing data more accessible to researchers, for example through data sharing agreements or by boosting data literacy.

Our results can help countries better understand if their statistical system is supporting the needs of researchers. Given that research is at the heart of evidence-based policy making, a statistical system oriented around making data available and easily accessible for these purposes is important. Countries could use the results to facilitate an understanding of the maturity of their statistical systems in supporting researchers, point to other countries that a particular country could learn from as it seeks to improve, and track progress or assess return on investment for funding related to statistical capacity building.

Our results face some limitations. First, the s2orc database does not currently contain non-English language articles. While the dominance of the English language in academic publication has been well-documented (Altbach 2007), the database omits articles published in other languages. Second, because of computational limitations, only 1 million articles were analyzed out of a sampling frame of around 10 million articles. Future work could bring in these additional articles to improve precision. Third, while the natural language processing model was able to classify whether or not the articles made use of data with 87% accuracy, the model could be extended in ways to extract more information from the articles, including more details about the subject of study and which particular data sources may have been used (surveys, censuses, administrative data, geospatial information etc.).

# References

Altbach, Philip G. “The imperial tongue: English as the dominating academic language.” *Economic and political Weekly* (2007): 3608-3611.

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.

Cameron, Grant J., Hai‐Anh H. Dang, Mustafa Dinc, James Foster, and Michael M. Lokshin. 2021. "Measuring the Statistical Capacity of Nations." *Oxford Bulletin of Economics and Statistics* 83(4): 870-896.

Courtioux, Pierre, François Métivier, and Antoine Rebérioux. 2022. "Nations Ranking in Scientific Competition: Countries Get What They Paid For." *Economic Modelling* 116: 105976.

Dang, Hai-Anh H, John Pullinger, Umar Serajuddin, and Brian Stacy. 2023. “Statistical Performance Indicators and Index: A New Tool to Measure Country Statistical Capacity.” *Scientific Data* 10(1): 146.

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.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini. 2021. “How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities.” *American Economic Review* 111 (5): 1442-80.

Jolliffe, Dean, Daniel Gerszon Mahler, Malarvizhi Veerappan, Talip Kilic, Philip Wollburg. 2023. “What Makes Public Sector Data Relevant for Development?” *World Bank Research Observer* 38 (2): 325-346.

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>.

Phillips, Brian J., and Kevin T. Greene. 2022. "Where is Conflict Research? Western Bias in the Literature on Armed Violence." *International Studies Review* 24 (3): viac038.

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.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” *arXiv preprint arXiv:1910.01108* (2019).

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.

Smirnov, Anatoly A., and Irina V. Stukova. "Determinants of integration approach in the agrarian sphere development in contexts of transformation." *Review of European studies* 7, no. 8 (2015): 8.

World Bank. 2021. “World Development Report 2021: Data for Better Lives.” Washington DC: World Bank. <https://wdr2021.worldbank.org/>

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

# Appendix

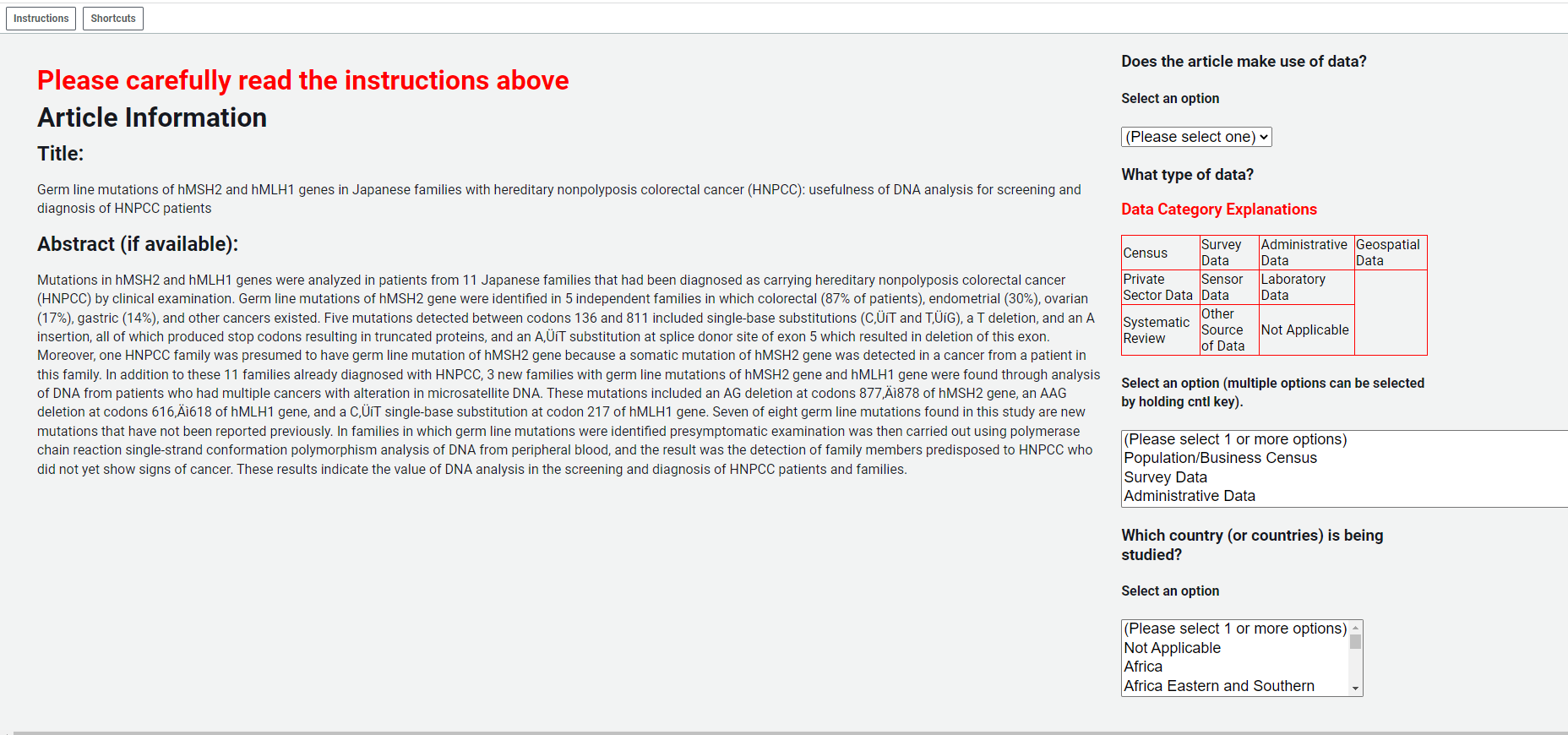
Countries shaded in dark orange have the lowest numbers of data use articles, countries in dark green have the highest Countries are grouped into five groups:

* **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 A.1. Country Production of Articles Using Data. Per capita and Total. 2017-19.*

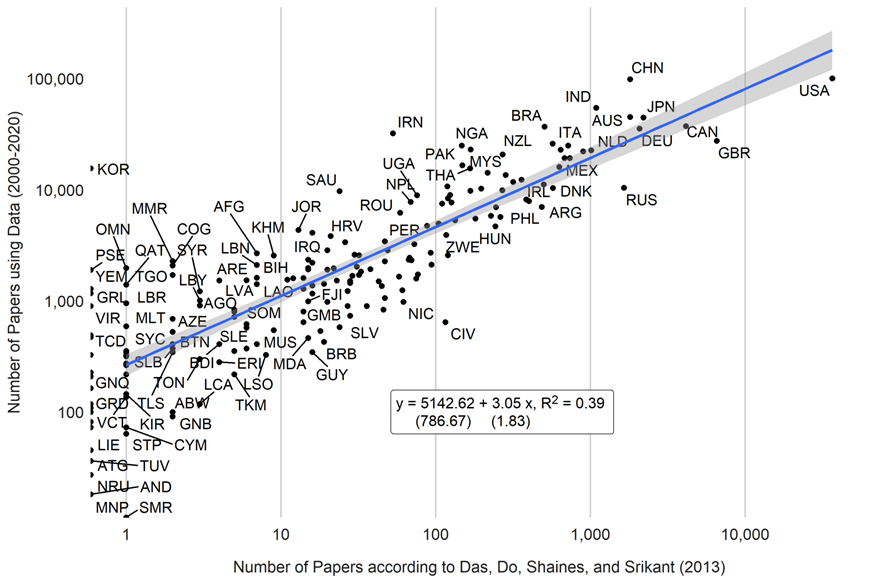
| Country | Articles using data per million persons | Total Number of Articles using data |
| --- | --- | --- |
| New Zealand | 35.55 | 133 |
| Norway | 22.56 | 98 |
| Ireland | 19.19 | 68 |
| Denmark | 17.89 | 89 |
| Sweden | 17.35 | 129 |
| Australia | 16.55 | 337 |
| Finland | 16.00 | 68 |
| Singapore | 15.55 | 86 |
| Estonia | 14.82 | 17 |
| Switzerland | 11.78 | 88 |
| Netherlands | 11.45 | 183 |
| Hong Kong SAR, China | 10.52 | 78 |
| Slovenia | 10.22 | 19 |
| Cyprus | 9.22 | 8 |
| Canada | 8.83 | 289 |
| Georgia | 7.88 | 27 |
| Latvia | 7.84 | 12 |
| Croatia | 7.63 | 26 |
| Lithuania | 7.52 | 15 |
| Israel | 7.14 | 52 |
| Malaysia | 7.10 | 200 |
| Greece | 7.03 | 67 |
| Portugal | 6.68 | 60 |
| Qatar | 6.65 | 20 |
| Botswana | 6.40 | 14 |
| Puerto Rico | 6.16 | 19 |
| Austria | 5.97 | 43 |
| Kosovo | 5.96 | 9 |
| North Macedonia | 5.62 | 7 |
| Mongolia | 5.16 | 14 |
| Eswatini | 5.13 | 4 |
| Belgium | 5.08 | 60 |
| Czechia | 4.87 | 43 |
| Oman | 4.78 | 17 |
| Spain | 4.69 | 211 |
| South Africa | 4.35 | 214 |
| Jordan | 4.14 | 49 |
| Italy | 4.10 | 238 |
| Slovak Republic | 4.03 | 10 |
| Bahrain | 4.02 | 4 |
| Germany | 3.80 | 273 |
| Jamaica | 3.79 | 5 |
| West Bank and Gaza | 3.77 | 18 |
| United Kingdom | 3.75 | 219 |
| Lebanon | 3.75 | 21 |
| Iran, Islamic Rep. | 3.71 | 266 |
| Kuwait | 3.68 | 12 |
| Serbia | 3.65 | 18 |
| Ghana | 3.60 | 110 |
| Hungary | 3.58 | 30 |
| Bosnia and Herzegovina | 3.57 | 10 |
| Saudi Arabia | 3.56 | 133 |
| Uruguay | 3.50 | 12 |
| Costa Rica | 3.41 | 16 |
| Korea, Rep. | 3.41 | 169 |
| Chile | 3.27 | 55 |
| Congo, Rep. | 3.23 | 15 |
| Romania | 3.15 | 47 |
| Poland | 3.14 | 134 |
| Bulgaria | 3.11 | 21 |
| Japan | 3.10 | 341 |
| Albania | 2.92 | 6 |
| Nepal | 2.91 | 66 |
| Kenya | 2.89 | 132 |
| Trinidad and Tobago | 2.85 | 2 |
| Panama | 2.84 | 10 |
| France | 2.72 | 165 |
| Timor-Leste | 2.60 | 3 |
| United States | 2.58 | 724 |
| Gambia, The | 2.52 | 4 |
| Sri Lanka | 2.42 | 42 |
| Armenia | 2.36 | 1 |
| Lao PDR | 2.22 | 19 |
| Turkiye | 2.18 | 157 |
| Uganda | 2.18 | 86 |
| Zimbabwe | 2.15 | 29 |
| Mauritius | 2.11 | 3 |
| Rwanda | 2.00 | 28 |
| Thailand | 1.96 | 134 |
| Tunisia | 1.94 | 21 |
| United Arab Emirates | 1.85 | 20 |
| Liberia | 1.74 | 9 |
| Brazil | 1.72 | 323 |
| Zambia | 1.72 | 33 |
| Ecuador | 1.63 | 19 |
| Moldova | 1.63 | 4 |
| Peru | 1.53 | 45 |
| Cambodia | 1.48 | 23 |
| Tanzania | 1.45 | 76 |
| Togo | 1.42 | 9 |
| Ethiopia | 1.41 | 177 |
| Nicaragua | 1.35 | 10 |
| Senegal | 1.33 | 22 |
| Colombia | 1.30 | 70 |
| Malawi | 1.27 | 21 |
| Papua New Guinea | 1.26 | 15 |
| Djibouti | 1.24 | 1 |
| Nigeria | 1.24 | 232 |
| Argentina | 1.23 | 57 |
| Indonesia | 1.23 | 339 |
| Gabon | 1.19 | 4 |
| Eritrea | 1.14 | 5 |
| Cuba | 1.12 | 12 |
| Libya | 1.12 | 2 |
| Benin | 1.11 | 11 |
| Mexico | 1.10 | 124 |
| Equatorial Guinea | 1.07 | 3 |
| Viet Nam | 1.07 | 101 |
| Haiti | 1.05 | 11 |
| Central African Republic | 1.02 | 3 |
| Niger | 0.98 | 26 |
| Bolivia | 0.93 | 12 |
| Guinea | 0.93 | 11 |
| China | 0.89 | 1,191 |
| Sierra Leone | 0.87 | 6 |
| Dominican Republic | 0.86 | 9 |
| Honduras | 0.84 | 10 |
| Ukraine | 0.84 | 33 |
| Azerbaijan | 0.83 | 9 |
| Bangladesh | 0.83 | 116 |
| Iraq | 0.83 | 32 |
| Kazakhstan | 0.81 | 13 |
| Cameroon | 0.79 | 24 |
| Pakistan | 0.77 | 177 |
| Lesotho | 0.75 | 2 |
| El Salvador | 0.74 | 4 |
| Egypt, Arab Rep. | 0.72 | 83 |
| Morocco | 0.72 | 27 |
| Russian Federation | 0.71 | 81 |
| Mauritania | 0.68 | 3 |
| Guatemala | 0.66 | 7 |
| Syrian Arab Republic | 0.66 | 16 |
| Myanmar | 0.63 | 40 |
| Kyrgyz Republic | 0.62 | 4 |
| Sudan | 0.61 | 26 |
| Belarus | 0.57 | 4 |
| Madagascar | 0.57 | 14 |
| Mozambique | 0.53 | 14 |
| Philippines | 0.52 | 52 |
| Afghanistan | 0.50 | 15 |
| Burkina Faso | 0.46 | 8 |
| South Sudan | 0.41 | 3 |
| Yemen, Rep. | 0.41 | 17 |
| Somalia | 0.40 | 6 |
| India | 0.36 | 466 |
| Venezuela, RB | 0.35 | 6 |
| Burundi | 0.34 | 4 |
| Mali | 0.34 | 8 |
| Algeria | 0.30 | 11 |
| Tajikistan | 0.25 | 2 |
| Chad | 0.23 | 3 |
| Congo, Dem. Rep. | 0.22 | 15 |
| Angola | 0.19 | 4 |
| Guinea-Bissau | 0.17 | 0 |
| Turkmenistan | 0.16 | 0 |
| Cote d'Ivoire | 0.14 | 3 |
| Korea, Dem. People's Rep. | 0.14 | 0 |
| Uzbekistan | 0.12 | 1 |

Note: Papers include all papers using data years 2017-2019. Population data comes from the World Bank. Only countries with a total population of more than 1 million persons shown.

*Figure A.1 Amazon Mturk Prompt*

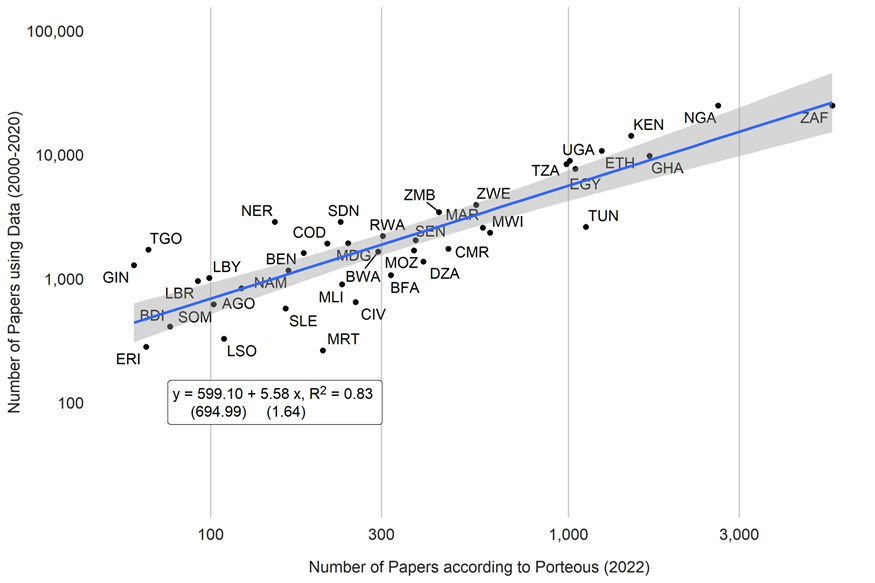
*Figure A.2. Comparison to Number of Academic Articles from other papers*

1. *Comparison with Das et al. (2013)*

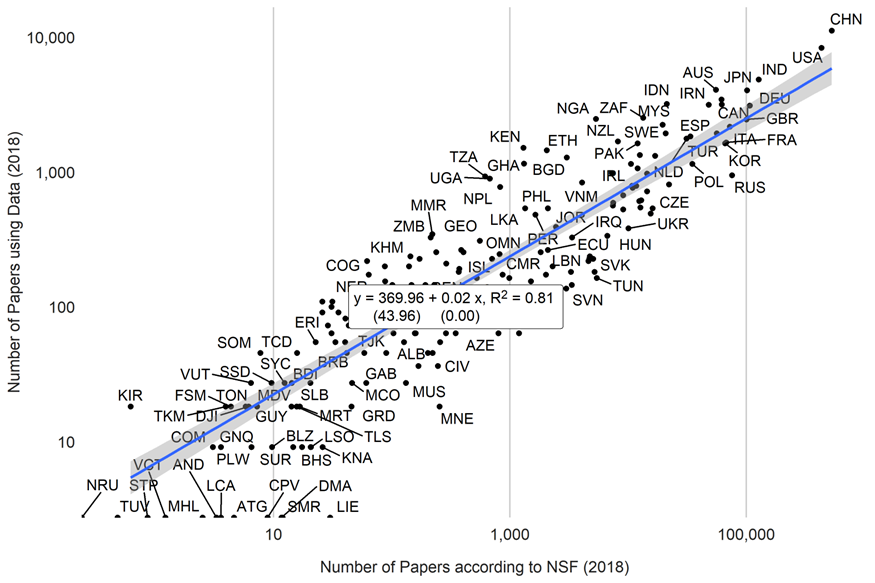


*Note*: Estimates from Das et al. (2013) are taken from Table A3 of their paper.

1. *Comparison with Porteous (2020)*

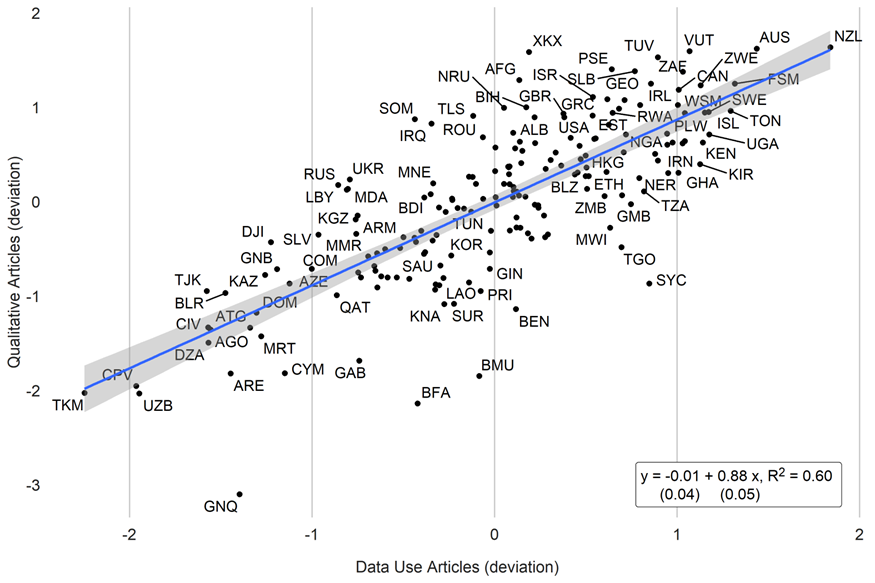
 *Note:* Estimates from Porteous (2020) are taken from Table 3 in that paper and include all journal publications from 2000 to 2019 (column 2 of Table 3).

1. *Comparison with NSF database of scientific and technical articles (2018)*

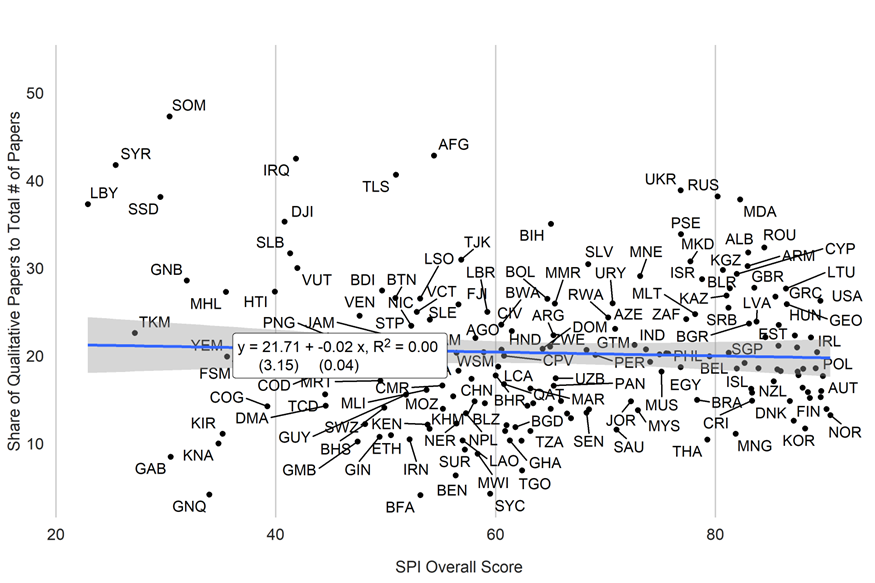


*Note*: Data from the NSF are papers produced for 2018.

*Figure A.3. Comparison of the Number of Qualitative and Quantitative Papers*

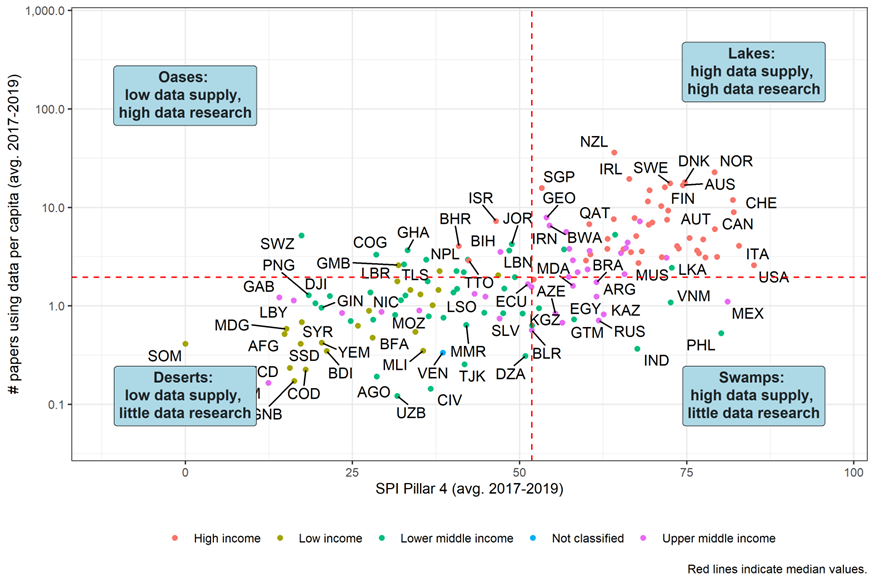


Note: The figure compares the residuals from regressing the number of qualitative or quantitative papers on GDP and population. The figure uses data from 2017-2019.

*Figure A.4. Share of Qualitative Papers vs SPI Overall Score. 2017-2019.* 

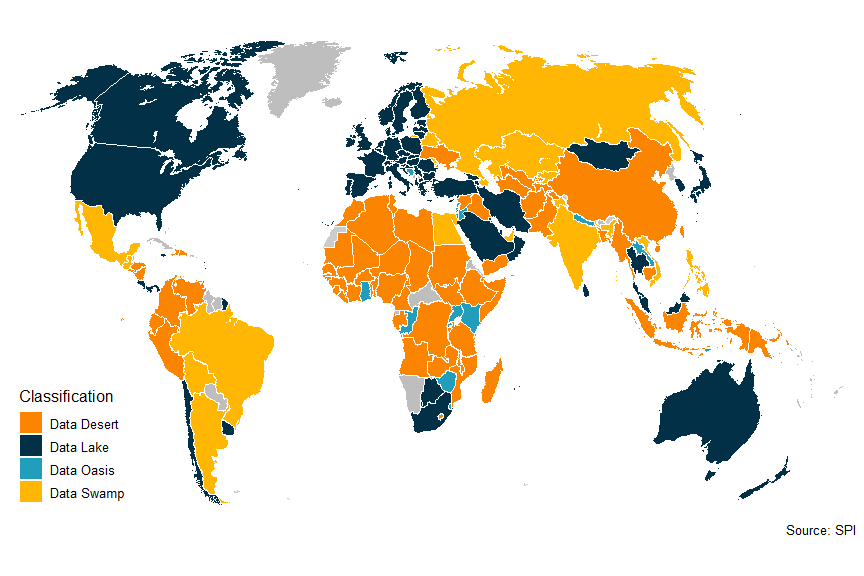
*Figure A.5. Relationship between Data Research and Data Supply*

1. *Country Scatterplot*



*Note:* Red lines indicate median values

1. *Map*

*Note*: a) On the y-axis, this figure plots the average number of data papers per capita between 2017 and 2019. SPI Pillar 4 on the x-axis tracks the availability of recent censuses, surveys, Civil Registration and Vital Statistics, and subnational data (for details see Dang et al. (2023). Red dashed lines represent median values.

*Table A.2. Relationship between Data Use in Academia and Data Sources from SPI.*

|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Population Census in past 10 years | 0.30+ |  |  |  |  |  |  |  |  |  |
|  | (0.16) |  |  |  |  |  |  |  |  |  |
| Agriculture Census in past 10 years |  | 0.16 |  |  |  |  |  |  |  |  |
|  |  | (0.12) |  |  |  |  |  |  |  |  |
| Business/Establishment Census in past 10 years |  |  | 0.12 |  |  |  |  |  |  |  |
|  |  |  | (0.11) |  |  |  |  |  |  |  |
| Household surveys: 2 or more in 10 years |  |  |  | 0.33 |  |  |  |  |  |  |
|  |  |  |  | (0.21) |  |  |  |  |  |  |
| Agriculture Surveys: 2 or more in 10 years |  |  |  |  | 0.22\* |  |  |  |  |  |
|  |  |  |  |  | (0.10) |  |  |  |  |  |
| Labor Force Surveys: 2 or more in 10 years |  |  |  |  |  | 0.36\* |  |  |  |  |
|  |  |  |  |  |  | (0.15) |  |  |  |  |
| Health Surveys: 2 or more in 10 years |  |  |  |  |  |  | 0.07 |  |  |  |
|  |  |  |  |  |  |  | (0.12) |  |  |  |
| Business/Establishments: 2 or more in 10 years |  |  |  |  |  |  |  | 0.20\* |  |  |
|  |  |  |  |  |  |  |  | (0.10) |  |  |
| Complete Civil Registration and Vital Statistics System |  |  |  |  |  |  |  |  | -0.21 |  |
|  |  |  |  |  |  |  |  |  | (0.18) |  |
| Availability of Data at 1st Admin Level (ODIN) Score |  |  |  |  |  |  |  |  |  | 1.27\*\* |
|  |  |  |  |  |  |  |  |  |  | (0.44) |
| Intercept | -4.45\*\*\* | -4.15\*\*\* | -4.14\*\*\* | -4.71\*\*\* | -4.37\*\*\* | -4.04\*\*\* | -4.26\*\*\* | -4.06\*\*\* | -4.73\*\*\* | -4.15\*\*\* |
|  | (1.16) | (1.18) | (1.21) | (1.25) | (1.16) | (1.16) | (1.19) | (1.19) | (1.26) | (1.14) |
| Log GDP per capita | 0.18\*\* | 0.18\*\* | 0.20\*\* | 0.22\*\*\* | 0.20\*\* | 0.16\* | 0.21\*\* | 0.18\*\* | 0.28\*\* | 0.19\*\* |
|  | (0.07) | (0.07) | (0.06) | (0.07) | (0.06) | (0.06) | (0.06) | (0.07) | (0.09) | (0.06) |
| Log Population | 0.30\*\*\* | 0.30\*\*\* | 0.29\*\*\* | 0.30\*\*\* | 0.31\*\*\* | 0.29\*\*\* | 0.29\*\*\* | 0.30\*\*\* | 0.29\*\*\* | 0.29\*\*\* |
|  | (0.05) | (0.05) | (0.06) | (0.05) | (0.05) | (0.05) | (0.06) | (0.05) | (0.06) | (0.05) |
| Log Qualitative Articles | 0.68\*\*\* | 0.67\*\*\* | 0.69\*\*\* | 0.65\*\*\* | 0.64\*\*\* | 0.66\*\*\* | 0.69\*\*\* | 0.66\*\*\* | 0.69\*\*\* | 0.64\*\*\* |
|  | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) | (0.08) |
| Observations | 168 | 168 | 168 | 168 | 168 | 168 | 168 | 168 | 168 | 168 |
| R Sq. | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| Note: GDP and population data are retrieved from the World Bank's World Development Indicators while the data sources are from the SPI. Papers include all papers using data years 2017-2019. \*\*\*=0.001 level, \*\*=0.01 level, \*=0.05 level, +=0.1 level. | | | | | | | | | | |

1. Stacy, Wang, Mahler, and Serajuddin are with the World Bank’s Development Data Group. Lucas Kitzmüller completed the work while at the European Bank for Reconstruction and Development (EBRD). Corresponding author: Brian Stacy ([bstacy@worldbank.org](mailto:bstacy@worldbank.org)). We are grateful for comments from Dean Jolliffe, Jishnu Das, Olivier Dupriez, and Patrick Brock. We acknowledge financial support from a World Bank Research Support Grant. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. Views presented are those of the authors and not necessarily of the EBRD. [↑](#footnote-ref-2)
2. A stratified sample of articles for the training set was taken. 15% of the articles were medical, 15% psychology, 25% economics, 25% political science, 10% were tagged as business, and 10% sociology articles. [↑](#footnote-ref-3)
3. We ended up not using this in the paper. [↑](#footnote-ref-4)
4. The SCI does not contain as many indicators as the SPI (25 indicators versus 51 in the SPI), and not all the indicators are similar. 11 indicators in the SCI overlap with the SPI: National accounts base year, consumer price index, balance of payments manual in use, government finance accounting, population census, agriculture census, health surveys, household consumption/income survey, CRVS availability, and the IMF SDDS status of the country. Using these indicators, the extended SPI series contains 35 indicators with data back to 2004. These include five indicators covering data use, 1 covering data services, 16 covering data products, 8 covering data sources, and 4 covering data infrastructure. The correlation between the extended SPI and the official SPI overall score is around 0.90. [↑](#footnote-ref-5)