Replication package for Missing Evidence: Tracking Academic Data Use around the World

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

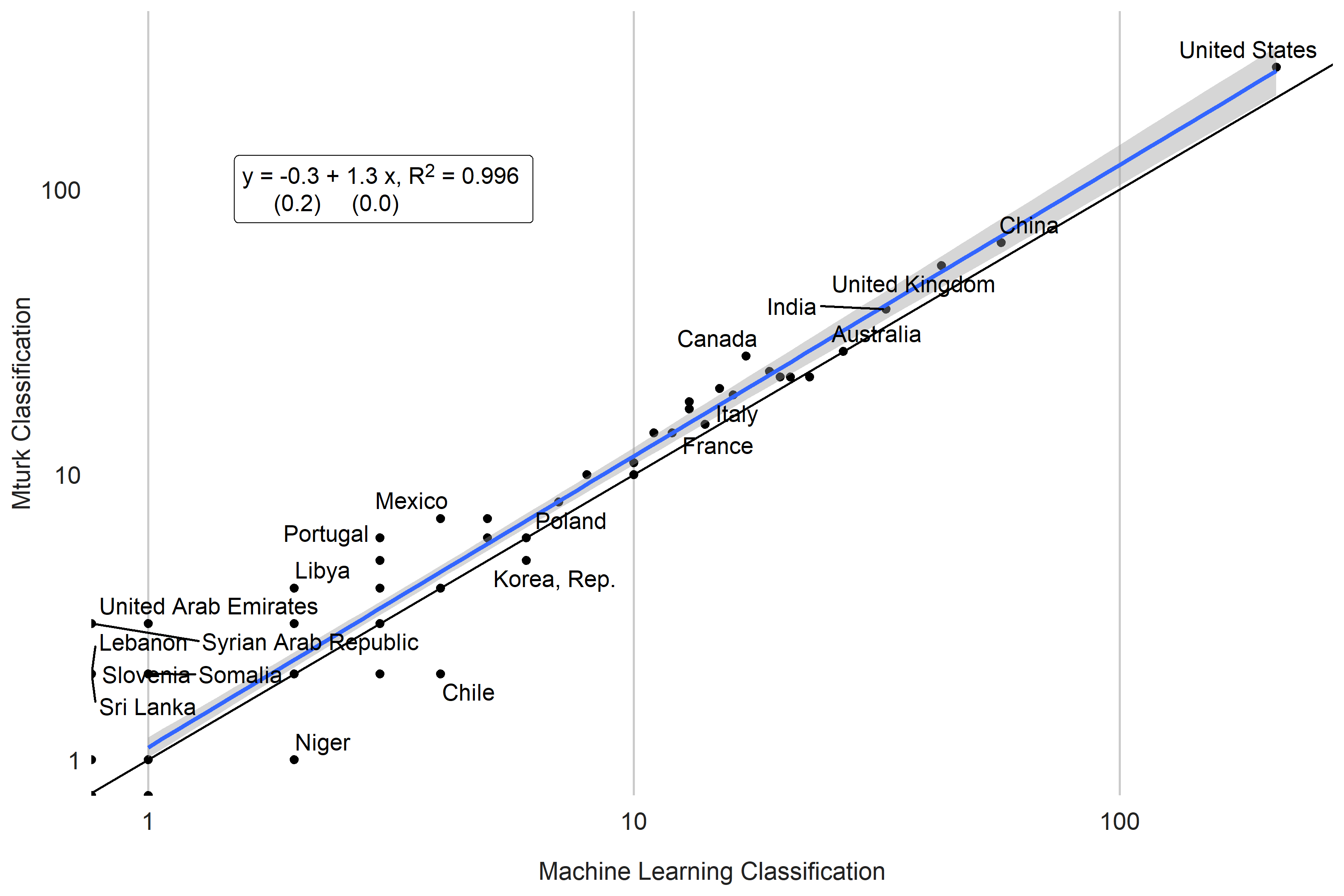
2023-11-13

# 1 Main Tables and Figures

## 1.1 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 | 0.44 | 0.64 |  | 28,571 | 2.8 |
| Economics | 0.21 | 0.68 |  | 62,241 | 6.0 |
| Medicine | 0.04 | 0.85 |  | 840,920 | 81.0 |
| Political Science | 0.58 | 0.33 |  | 26,185 | 2.5 |
| Psychology | 0.25 | 0.70 |  | 44,191 | 4.3 |
| Sociology | 0.10 | 0.33 |  | 35,640 | 3.4 |

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



## 1.3 Figure 2. Article Examples

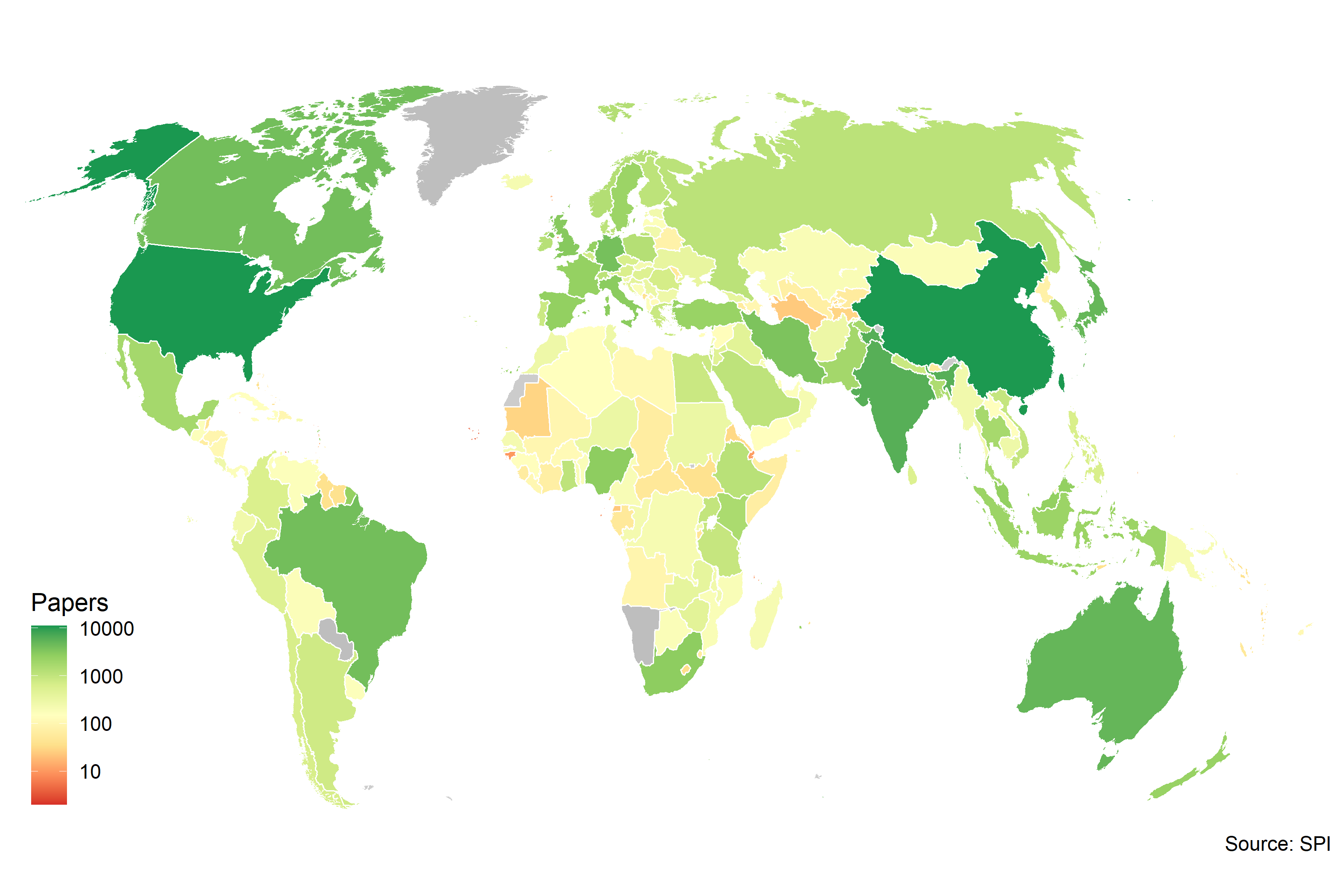
|  |
| --- |
| Article Example - No Data Use |

Article Example - No Data Use

|  |
| --- |
| Article Example - Data Use |

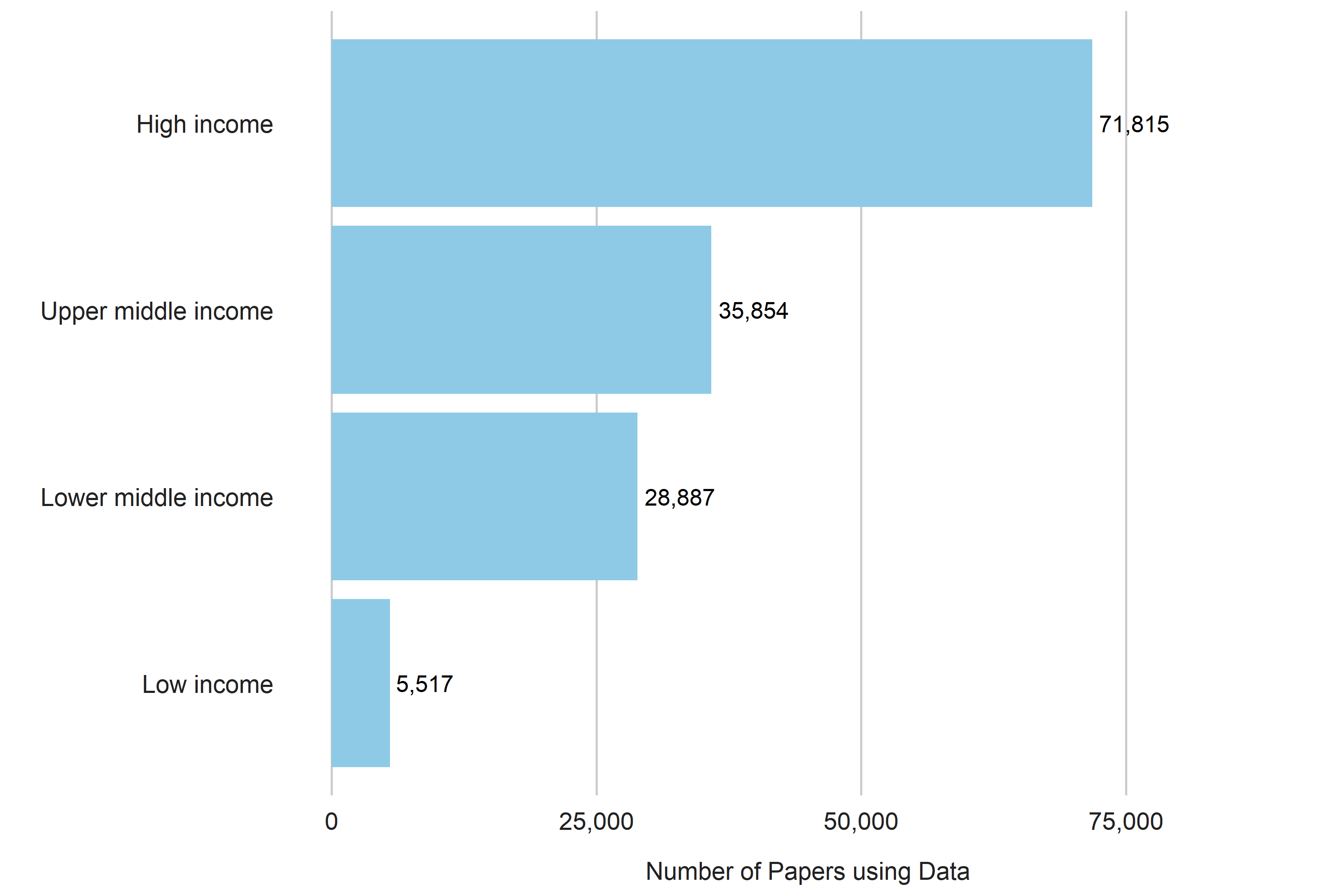
Article Example - Data Use

## 1.4 Figure 3. Number of Articles using Data by Country (2000-2020)

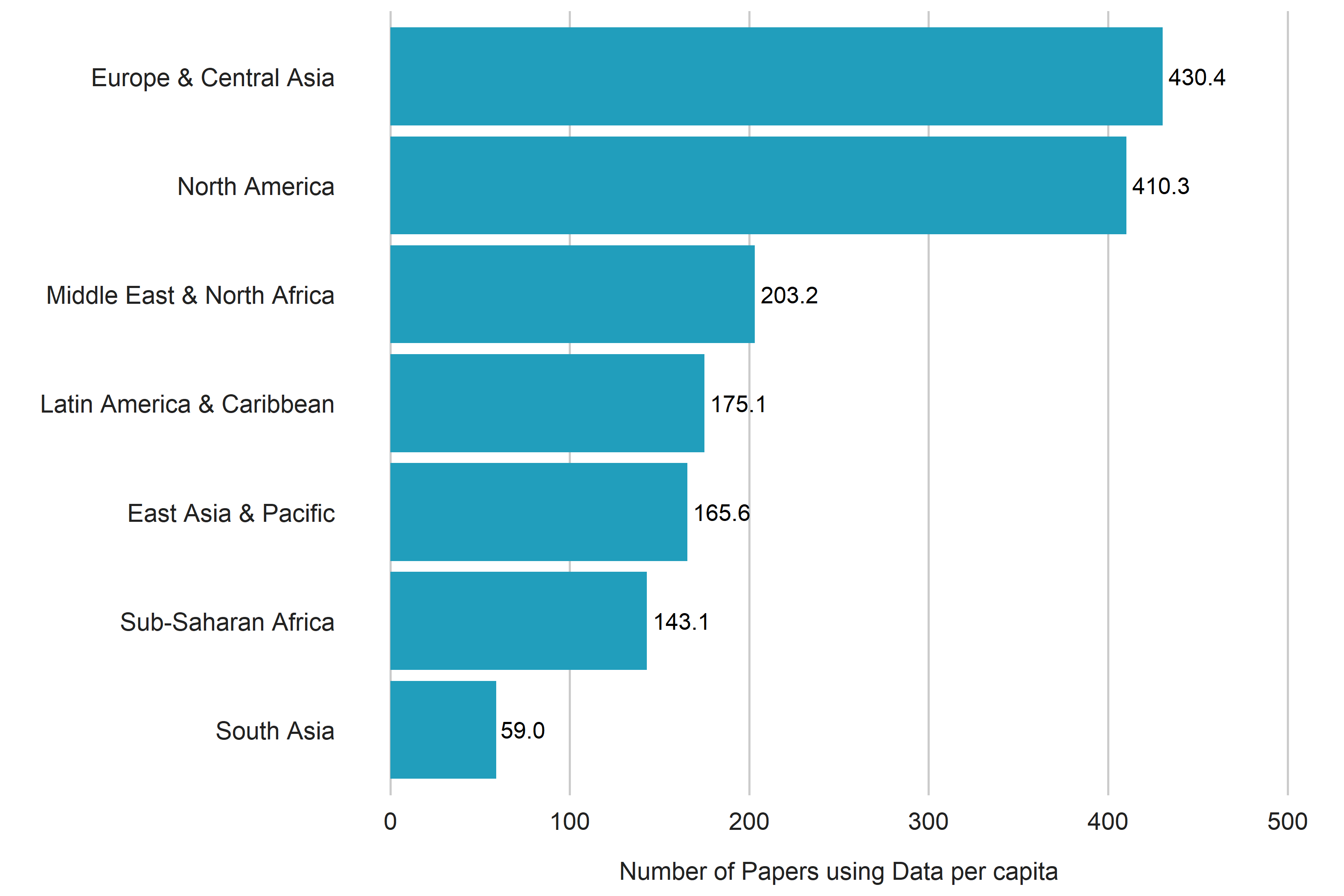


## 1.5 Figure 4. Number of Articles Using Data by Income Group and Region

By Income Group

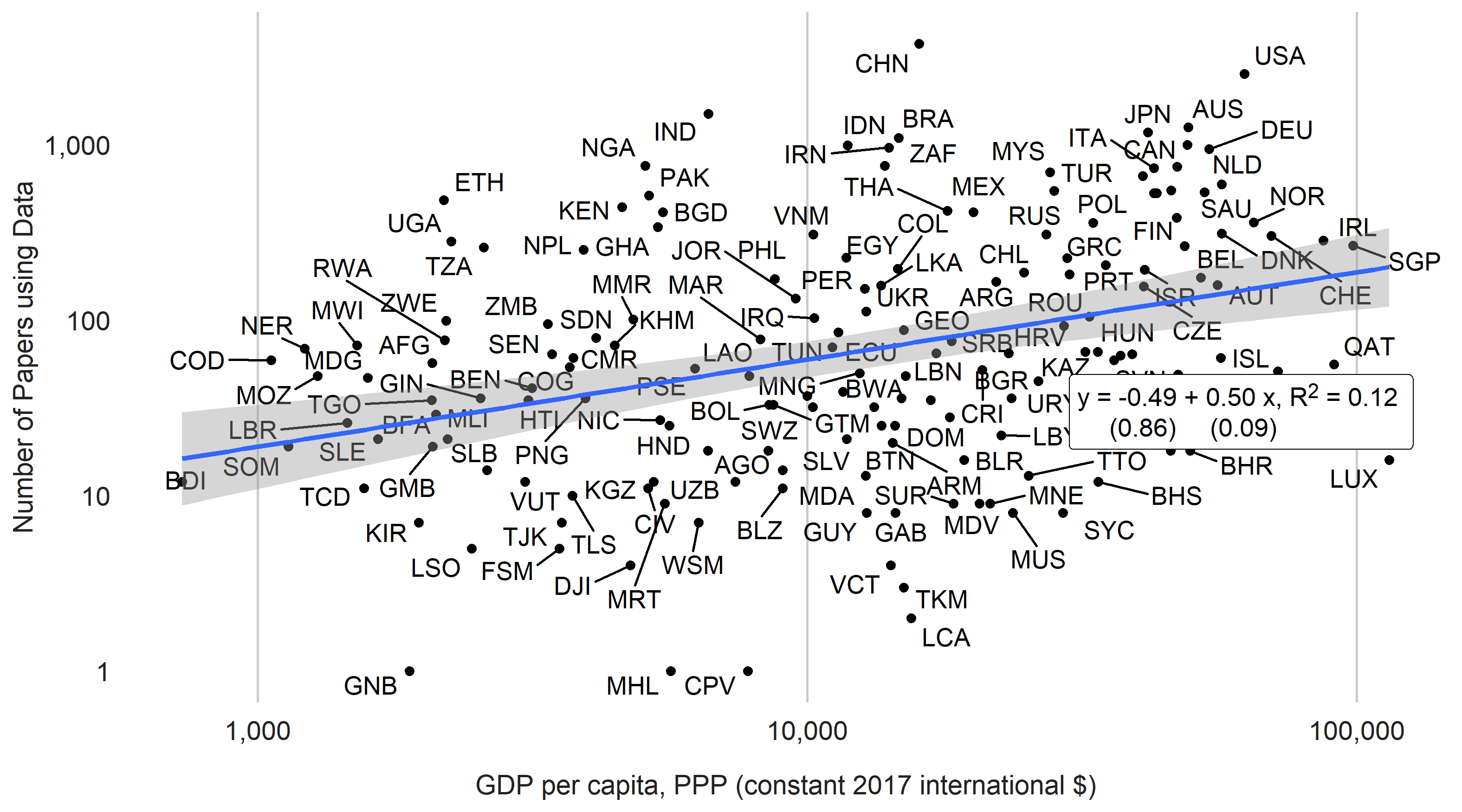


By region per million persons

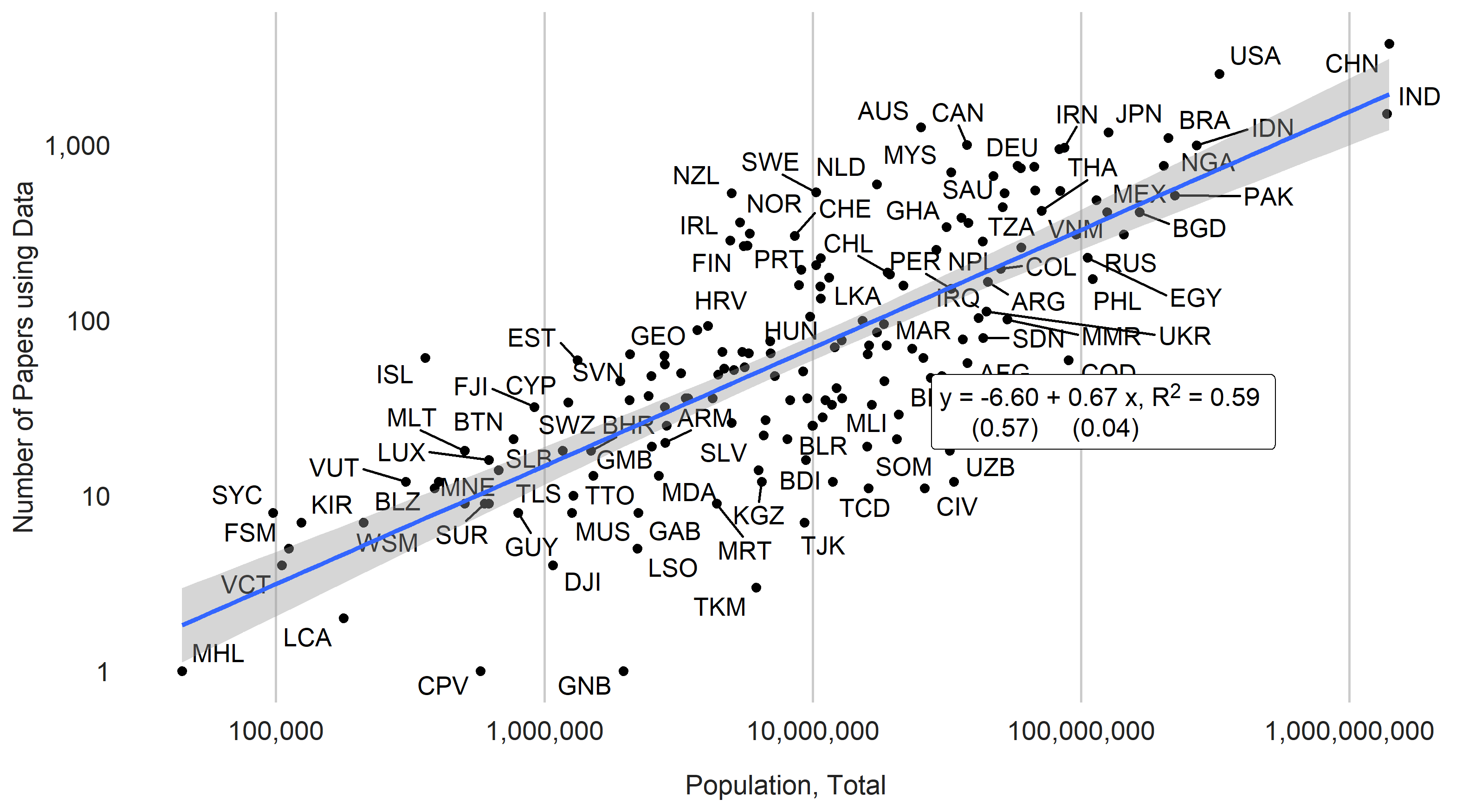


## 1.6 Figure 5. Relationship between Papers using Data and Development Outcomes

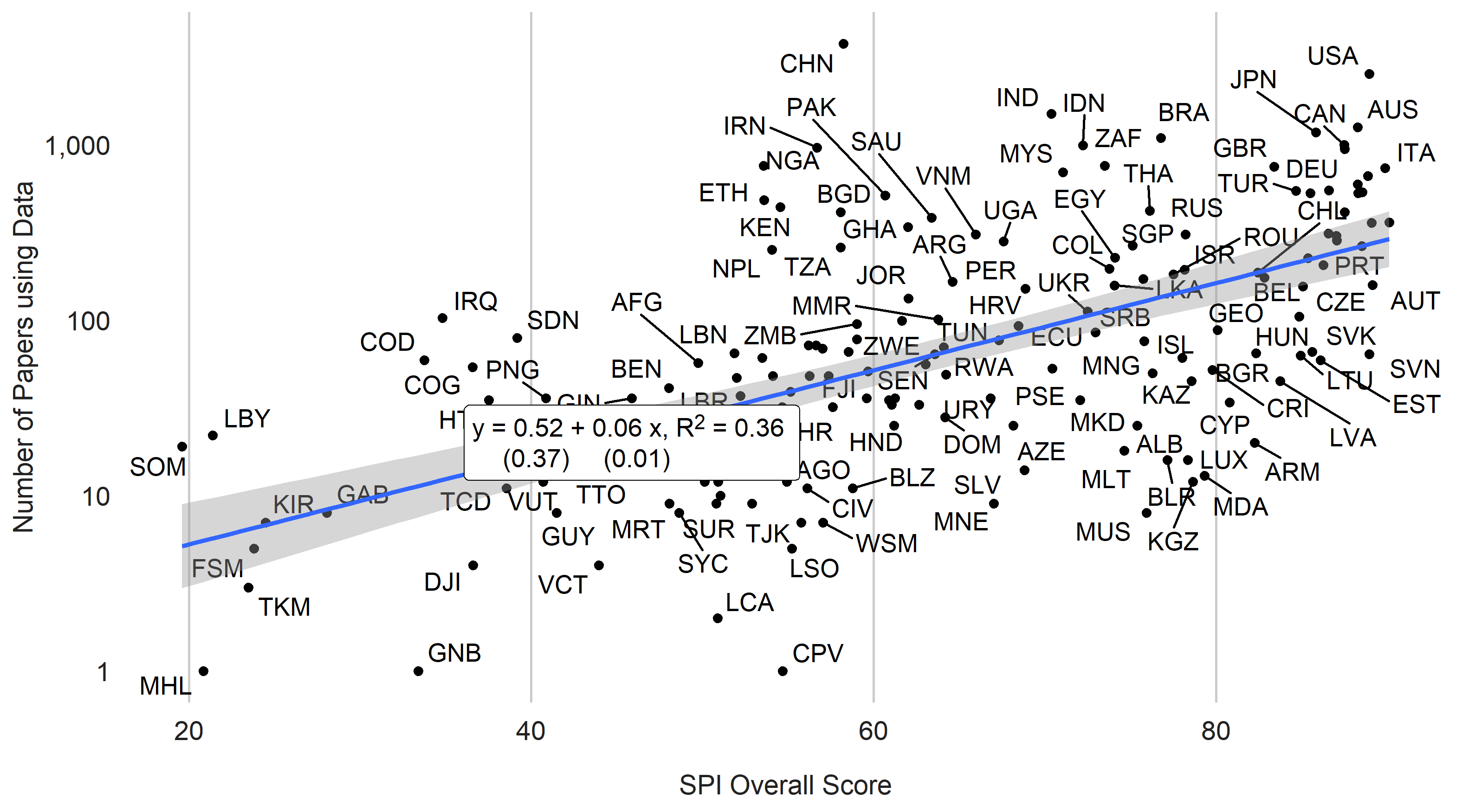
GDP per capita



Population



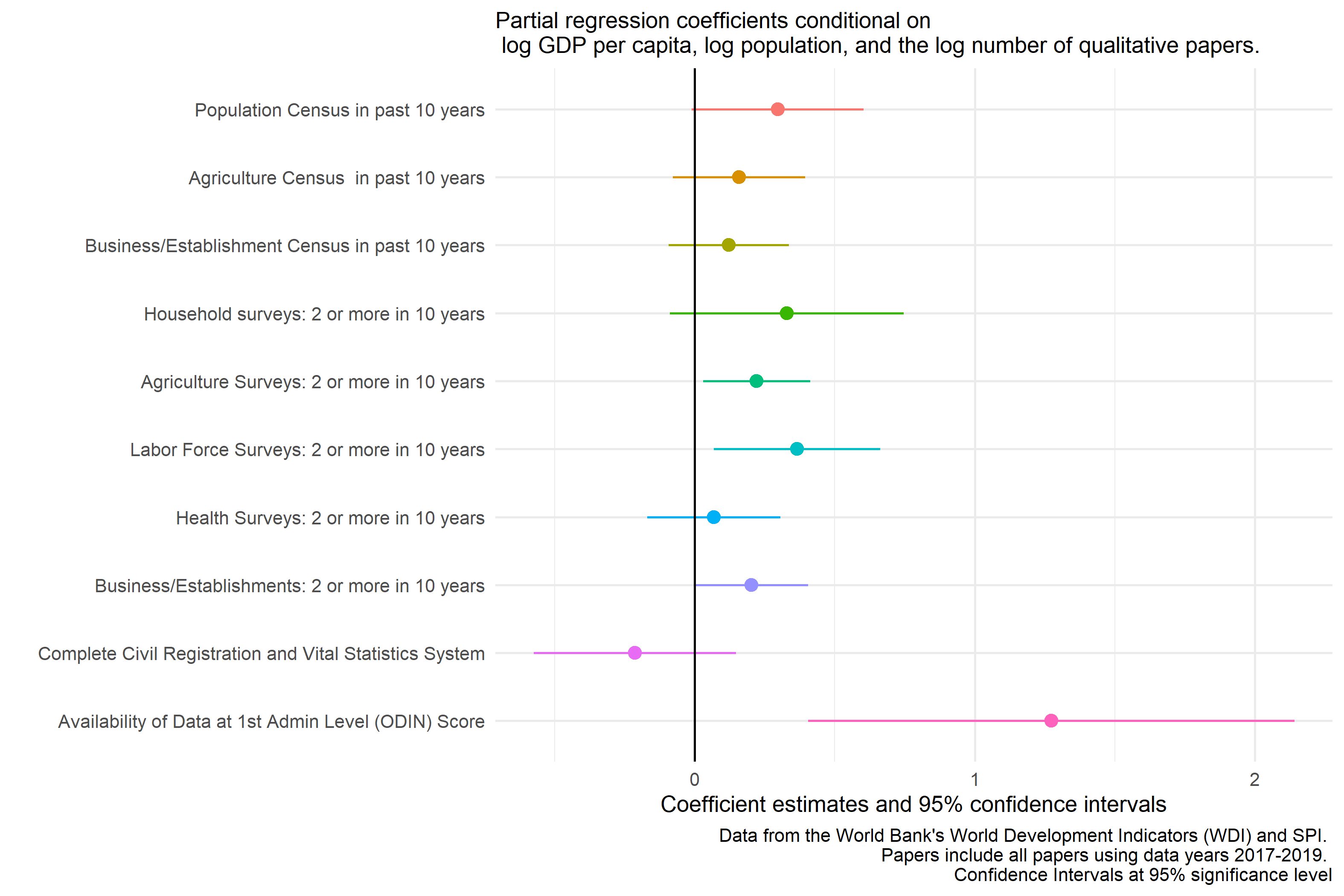
SPI Overall Score



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

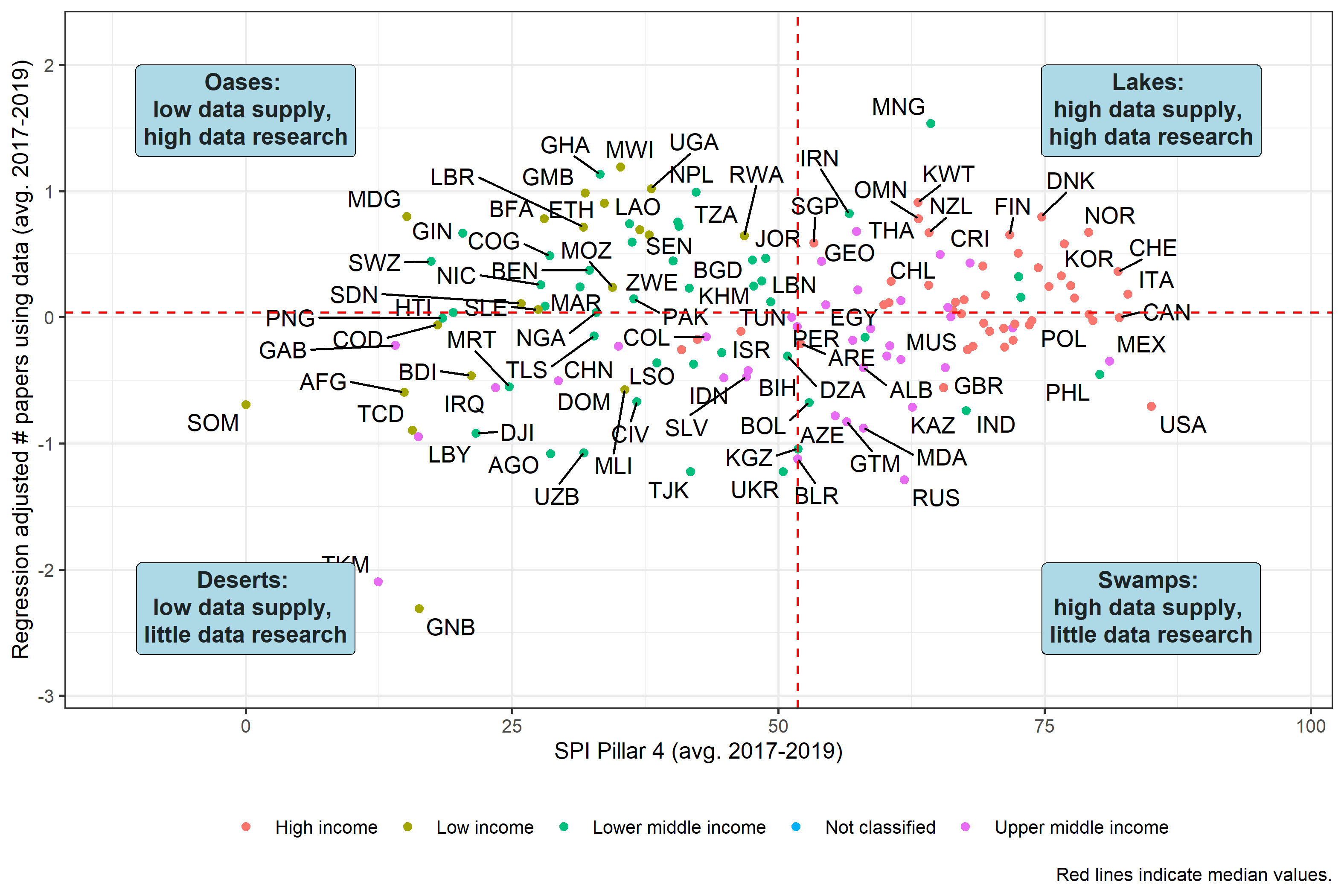
|  | (1) | (2) | (3) | (4) |
| --- | --- | --- | --- | --- |
| (Intercept) | -12.41\*\*\* | -10.66\*\*\* | -3.96\*\*\* | -3.57\*\* |
|  | (0.82) | (0.95) | (1.14) | (1.17) |
| Log GDP per capita | 0.58\*\*\* | 0.36\*\*\* | 0.12+ | 0.10 |
|  | (0.05) | (0.07) | (0.07) | (0.08) |
| Log Population | 0.70\*\*\* | 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(qual\_papers + 1) |  |  | 0.64\*\*\* | 0.65\*\*\* |
|  |  |  | (0.08) | (0.08) |
| SPI Data Sources Score |  |  |  | 0.01\* |
|  |  |  |  | (0.00) |
| Num.Obs. | 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 |
| Std.Errors | 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 | | | | |

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

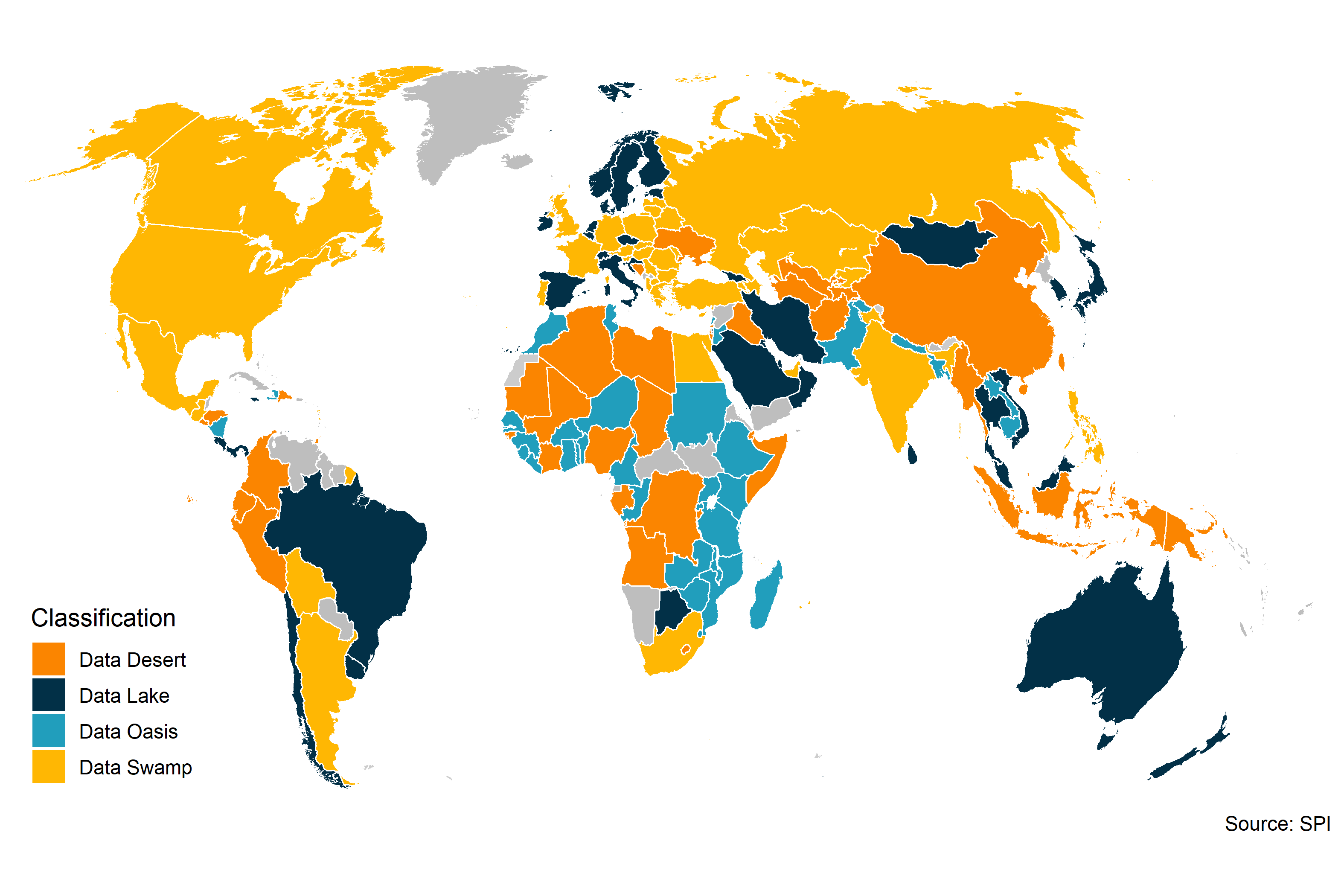


## 1.9 Figure 7. Relationship between Data Use and Data Supply

a)Country Scatterplot



1. Map



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

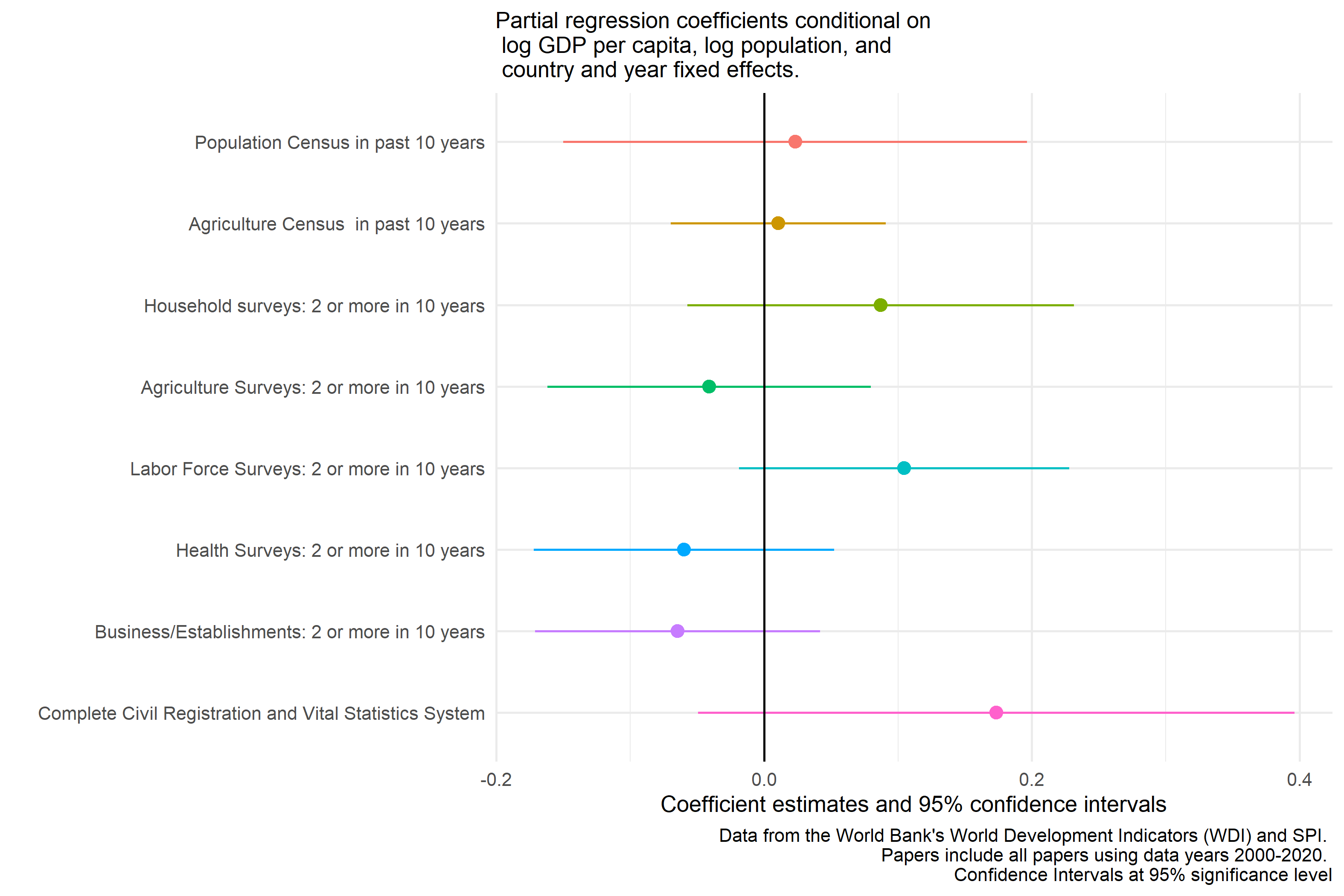
| group | Data Desert | Data Oasis | Data Swamp | Data Lake | Number of countries |
| --- | --- | --- | --- | --- | --- |
| East Asia & Pacific | 29% | 12% | 6% | 53% | 17 |
| Europe & Central Asia | 11% | 0% | 58% | 31% | 45 |
| Latin America & Caribbean | 37% | 11% | 21% | 32% | 19 |
| Middle East & 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 |
| High income | 7% | 0% | 38% | 56% | 45 |
| Low income | 35% | 65% | 0% | 0% | 20 |
| Lower middle income | 31% | 49% | 11% | 9% | 45 |
| Upper middle income | 33% | 0% | 44% | 22% | 36 |
| Overall | 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.

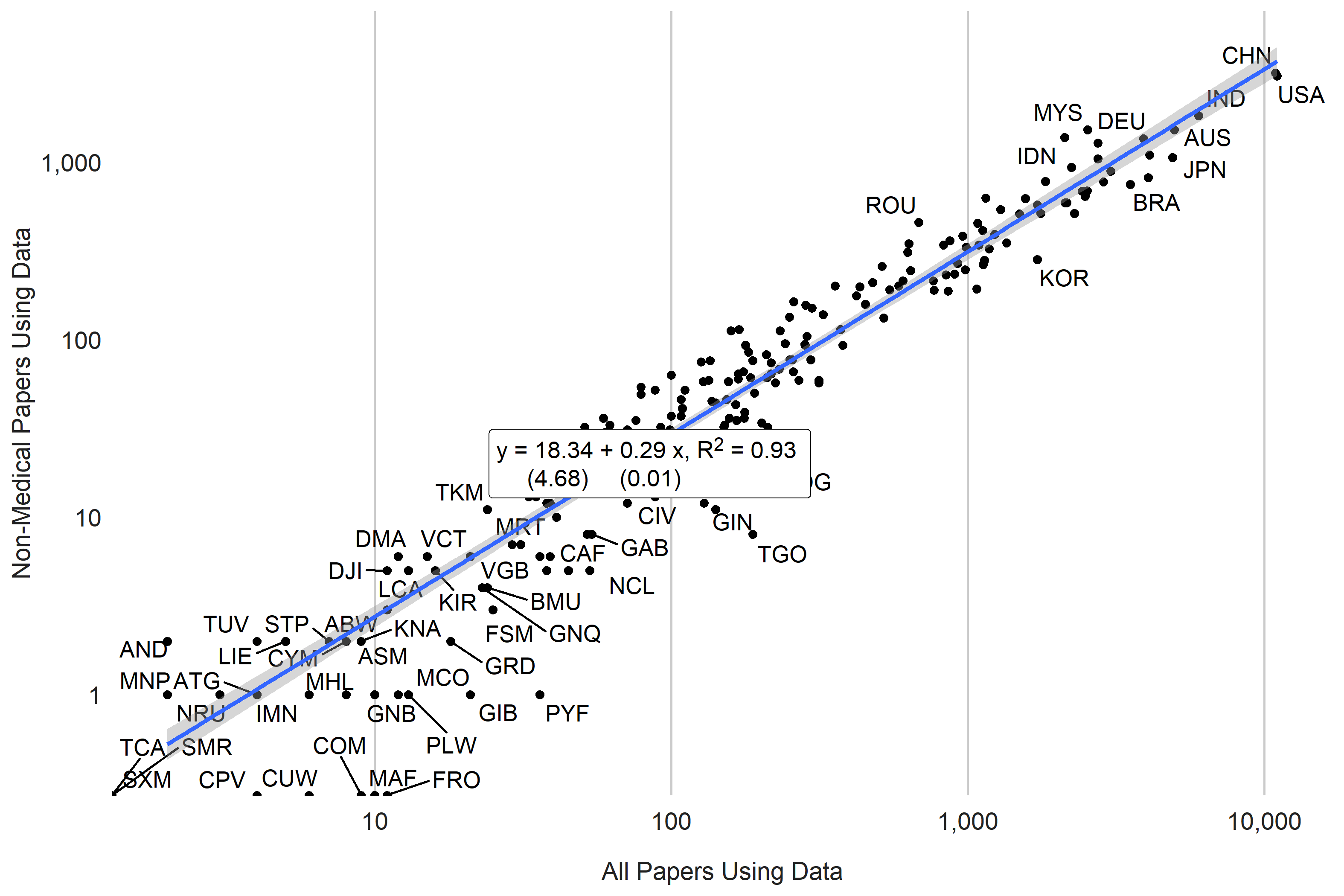
## 1.11 Table 4. Longitudinal Relationships between Number of Papers Using Data and Statistical Performance

|  | (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) |
| Num.Obs. | 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 |
| 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 | | | | | | |

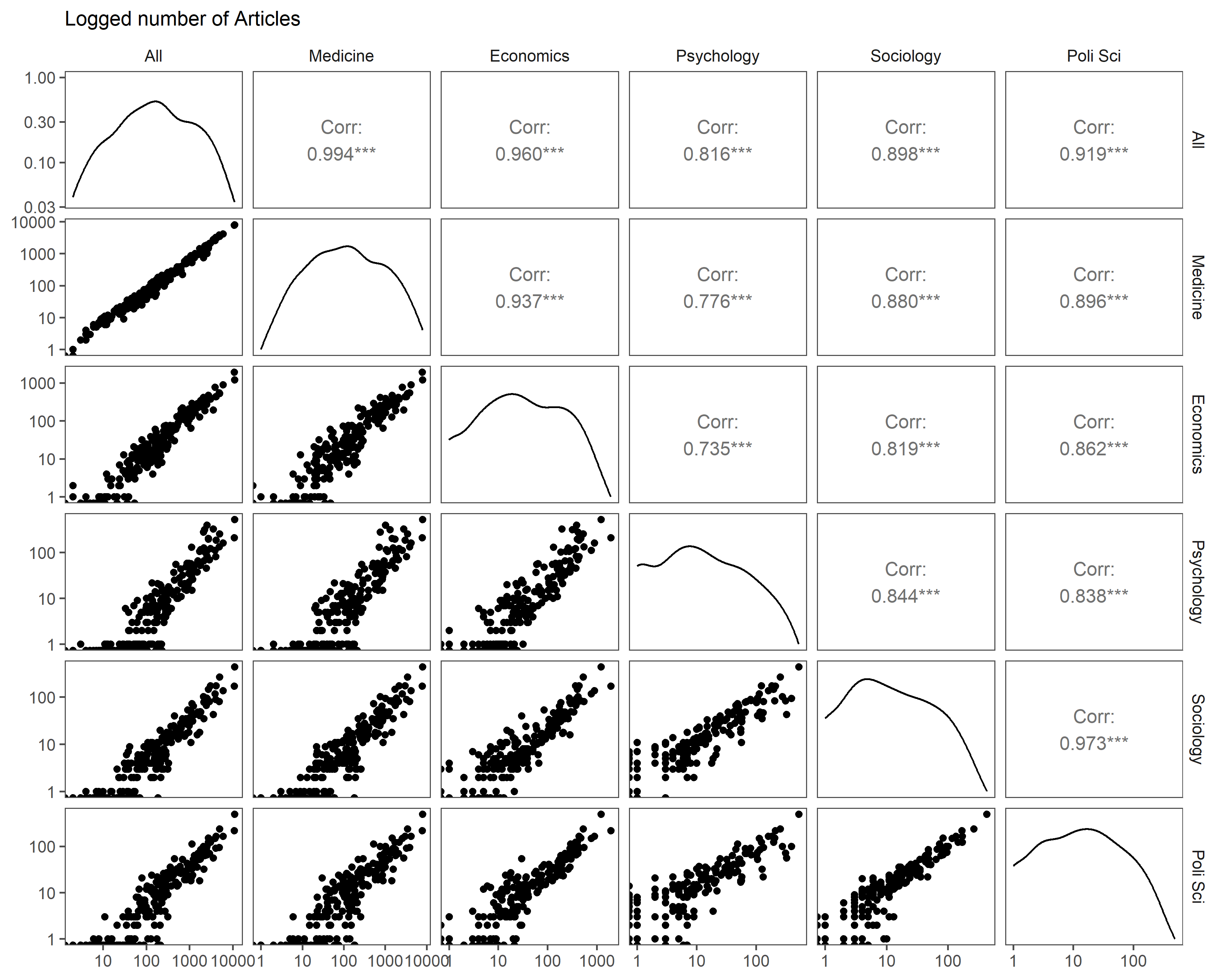
## 1.12 Figure 8. Panel Relationships Between Number of Papers Using Data and Data Sources



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



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



# 2 Appendix

## 2.1 Table A.1. Country Scores on Academic Data Use per capita. Year 2019.

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:

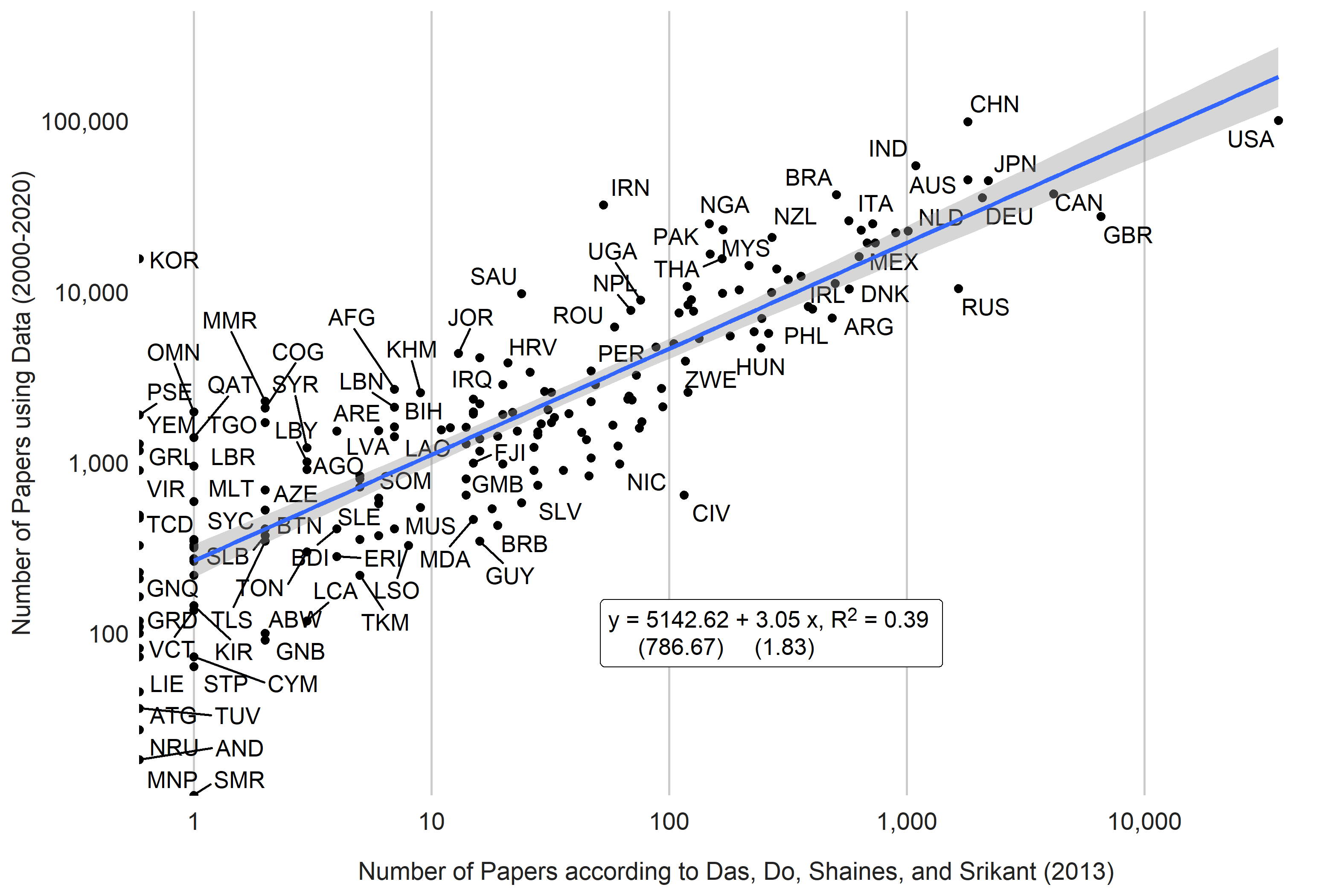
* **Top Quintile**: Countries in the Top quintile are classified in this group. Shading in dark green.
* **4th Quintile**: Countries in the 4th quintile, or those above the 60th percentile but below the 80th percentile are in this group. Shading in light green.
* **3rd Quintile**: Countries in the 3rd 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 .

| 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 |

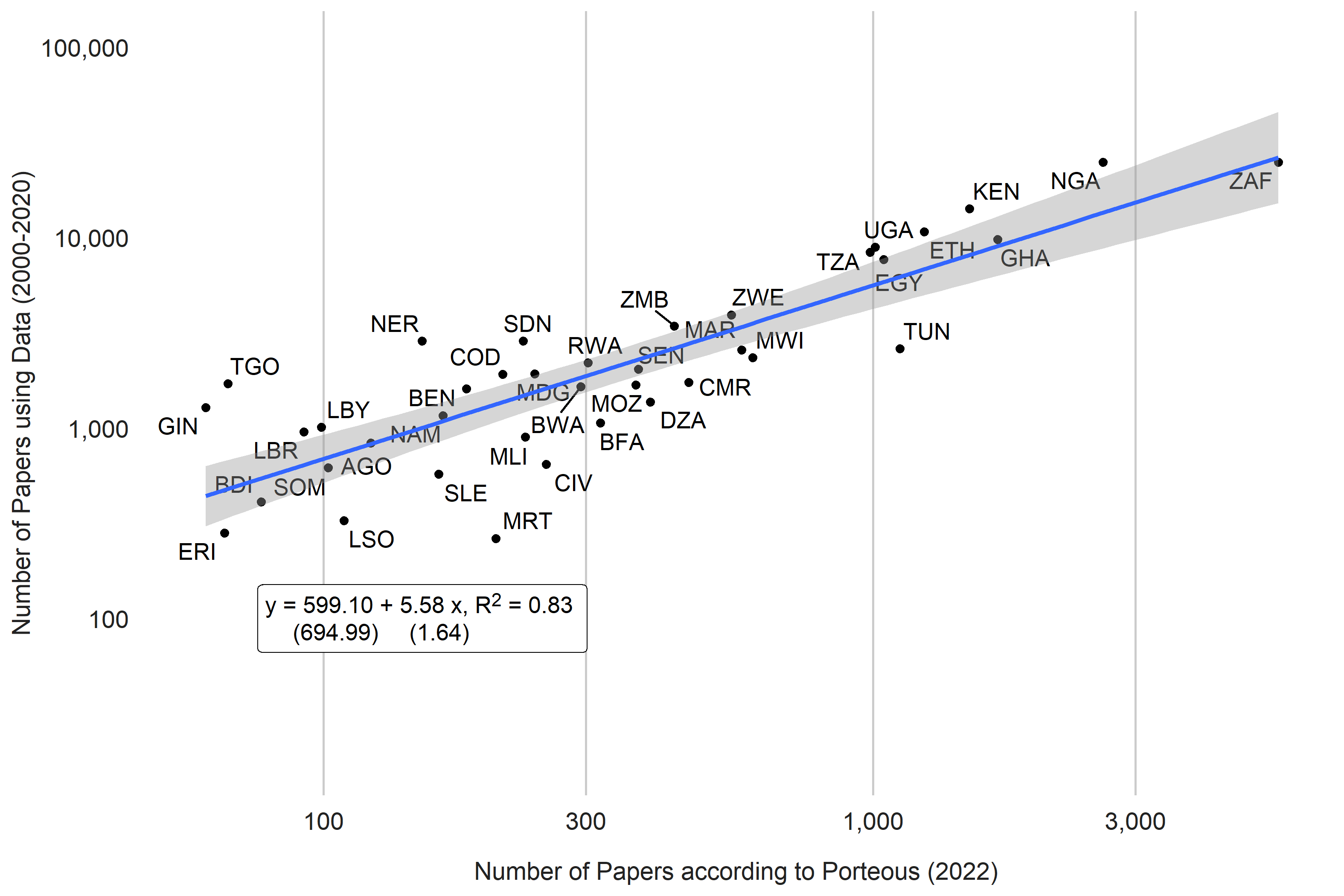
## 2.2 Figure A.1 Amazon Mturk Prompt

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

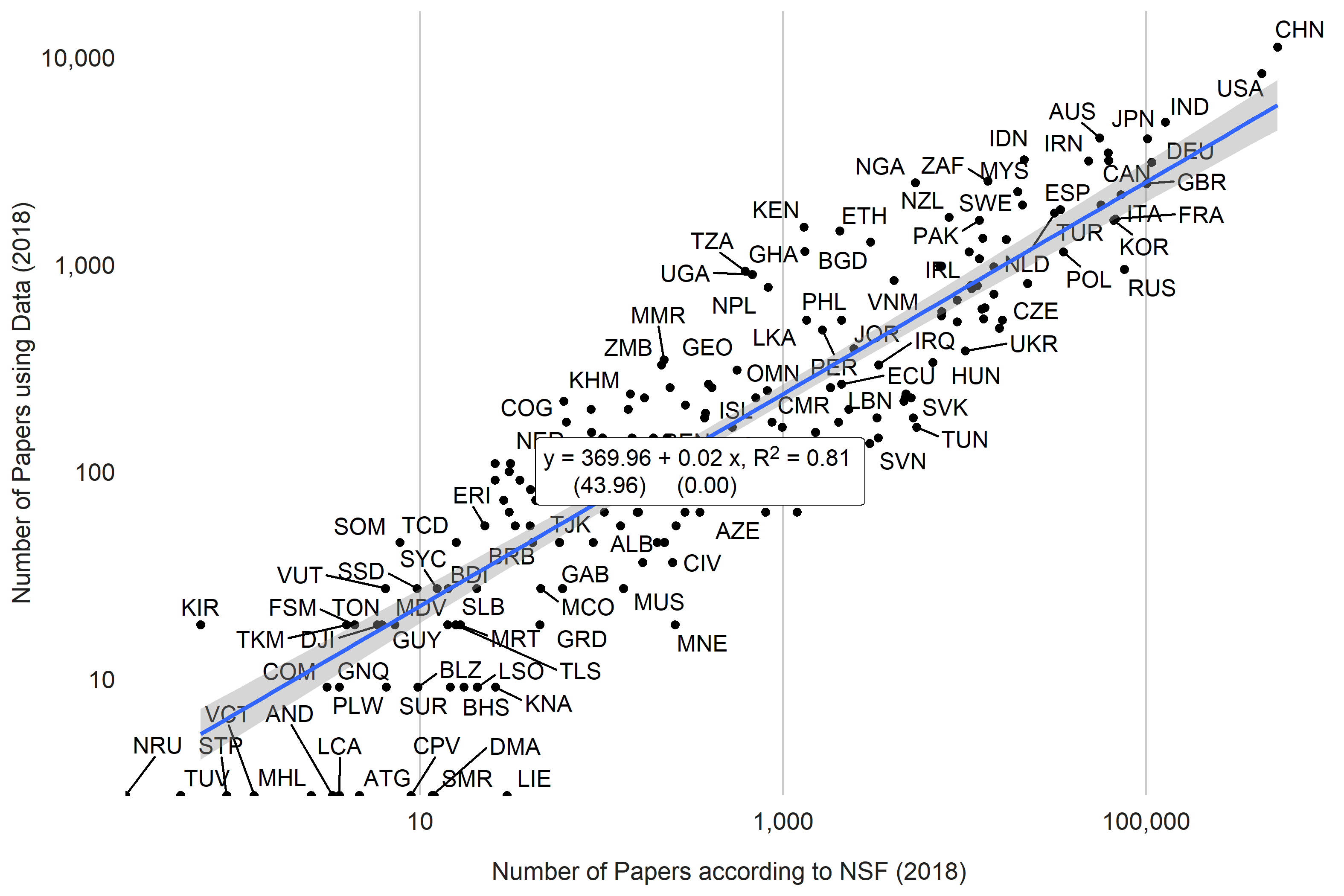
1. Comparison with Das et al. (2013)



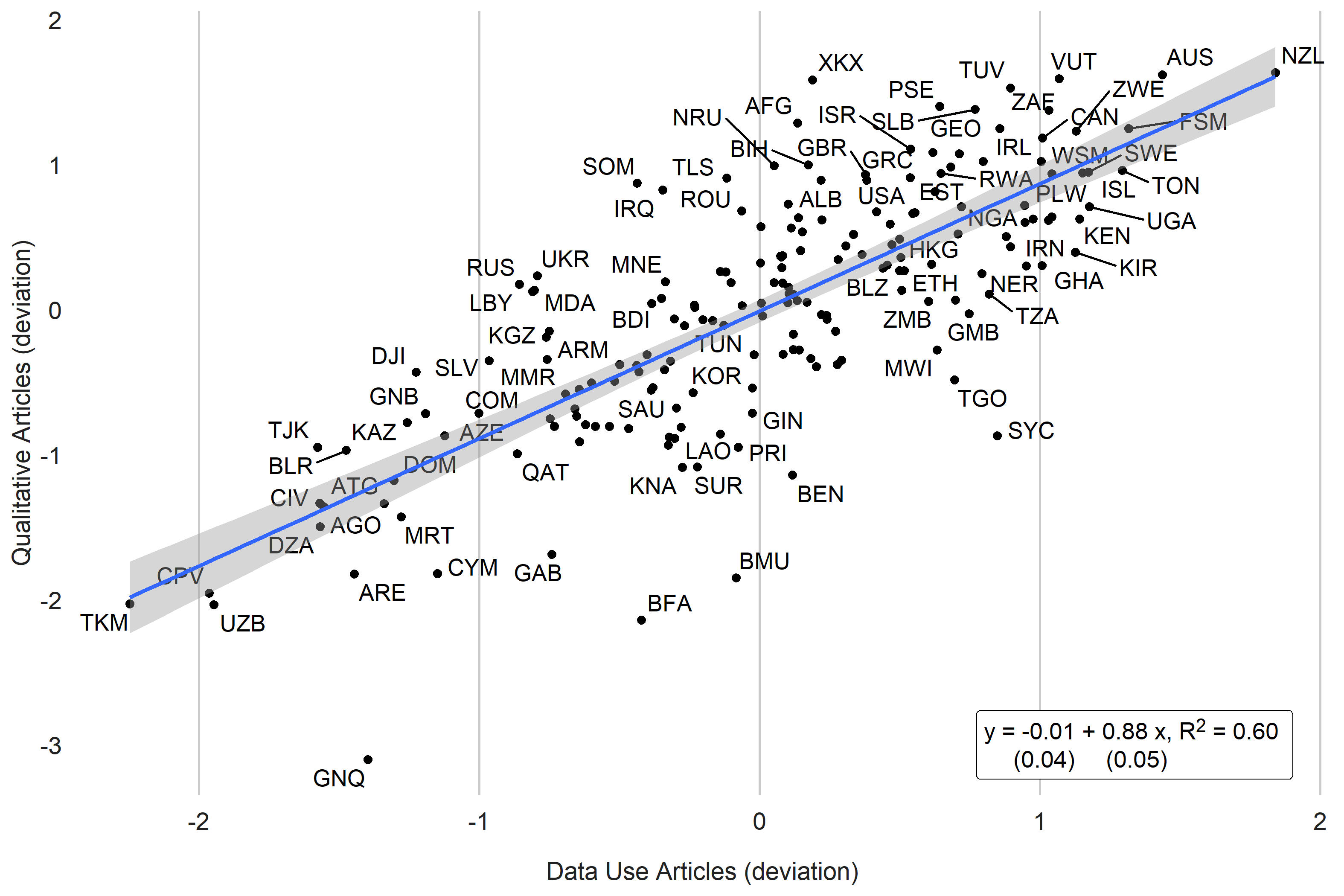
1. Comparison with Porteous (2020)



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



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



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

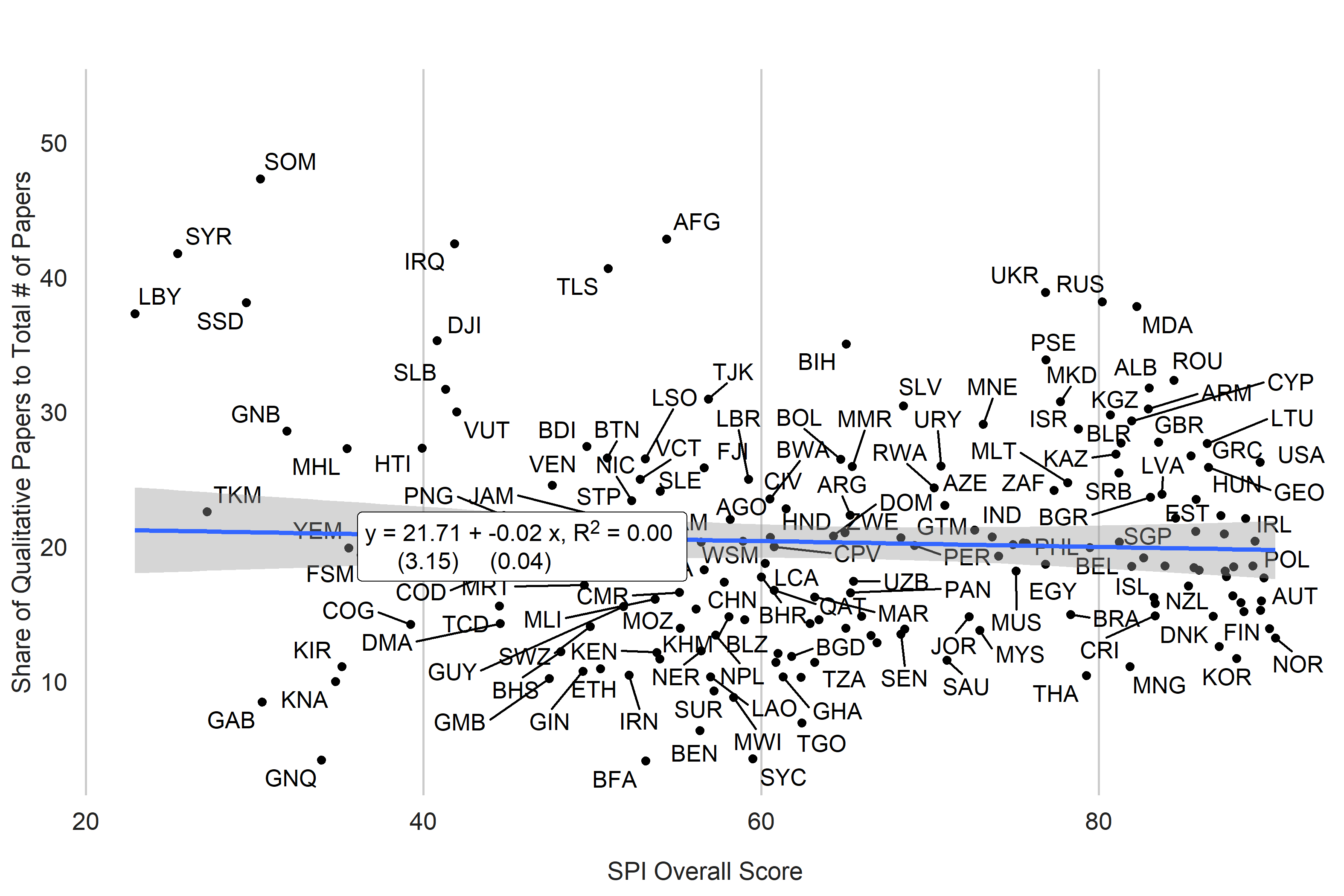
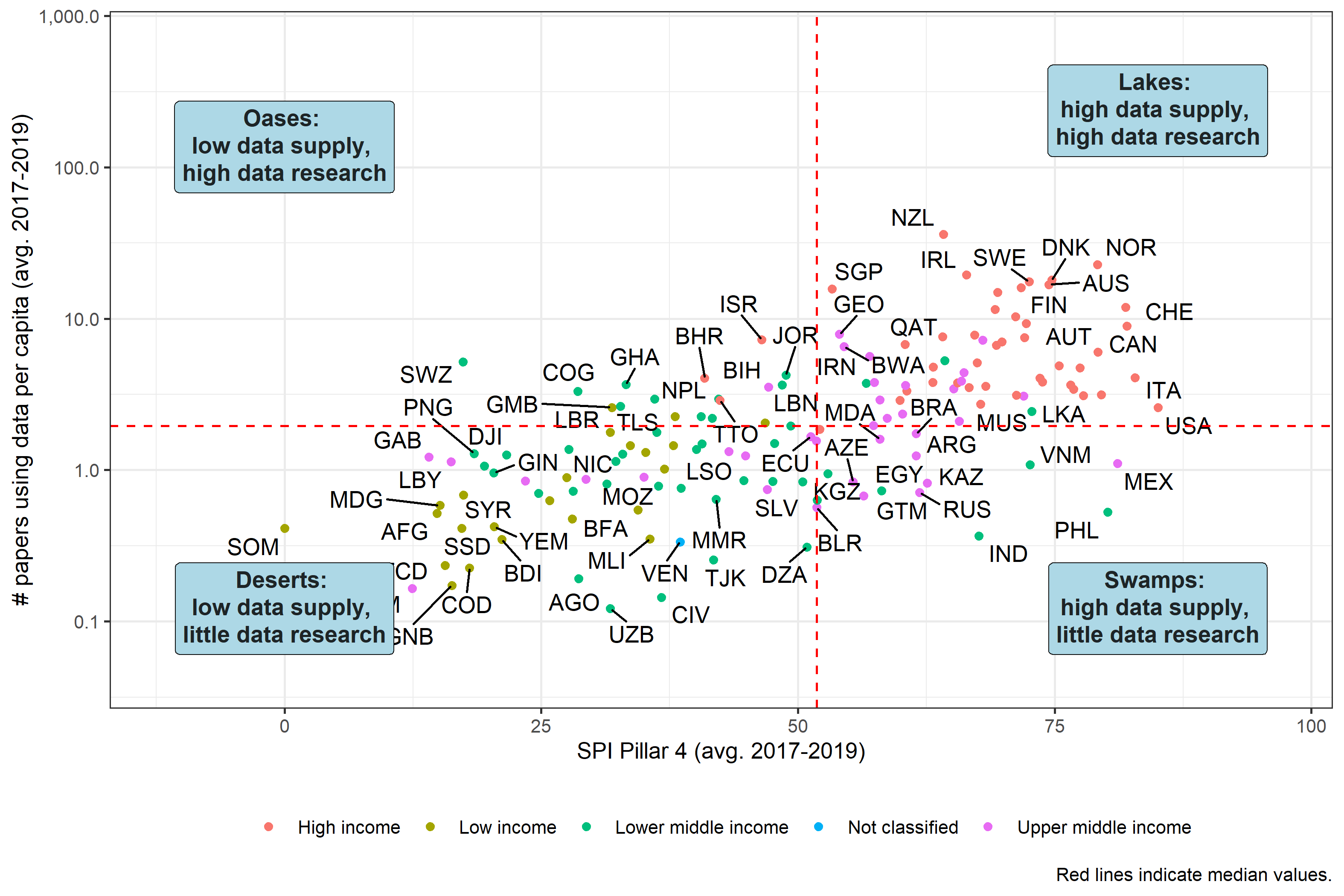
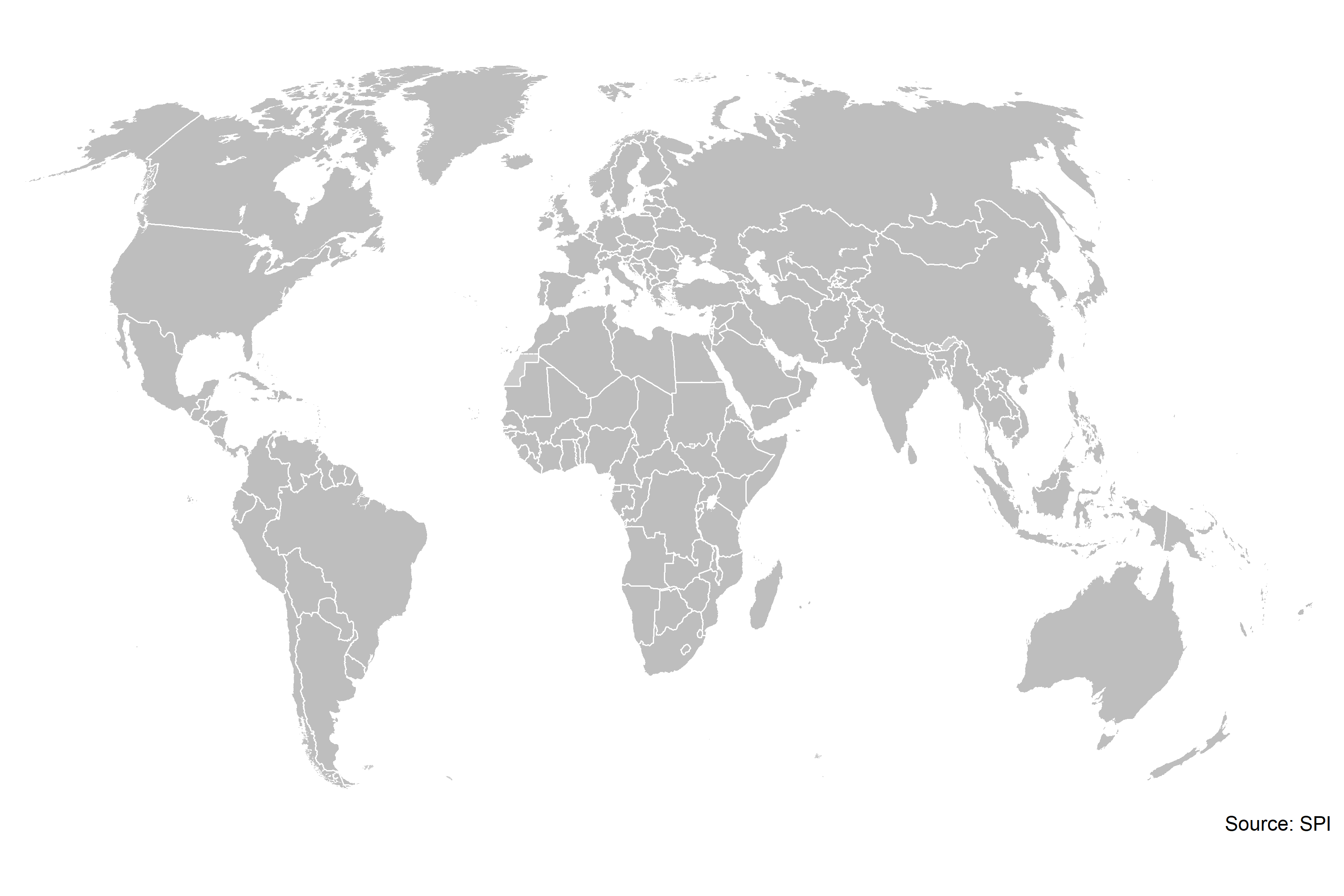


Figure A.5. Relationship between Data Research and Data Supply

1. Country Scatterplot



1. Map



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

|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (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(qual\_papers + 1) | 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) |
| 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) |
| N | 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 |
| Data from the World Bank's World Development Indicators (WDI) and SPI. Papers include all papers using data years 2017-2019.  \*\*\*=0.001 level  \*\*=0.01 level  \*=0.05 level  +=0.1 level | | | | | | | | | | |