Data Article Title

Pooled cross-sectional panel of the 2015-2018 PISA student questionnaire data files for the evaluation of youth related strategies implemented under the UN 2030 Agenda for sustainable development

Authors

Ibrahim Niankara[1], Rachidatou I. Traoret[2]

Affiliations

1. College of Business, Al Ain University, Abu Dhabi, UAE
2. Department of Economics, Université Aube Nouvelle, Ouagadougou, Burkina Faso

Corresponding author(s)

Ibrahim Niankara (ibrahim.niankara@aau.ac.ae)

Abstract

This data article is a follow up to the cross-national data on the environmental affection and cognition of adolescent students of varying levels of interest in ecosystem services and sustainability [1]. The data is being provided as a pooled cross-sectional panel of the 2015 [2] and 2018 [3] publicly released student questionnaire data files from the Programme for International Student assessment (PISA). The 2015 cross-section was used in the studies "Interest in the biosphere and students' environmental awareness and optimism: A global perspective" [4] and "Scientific media dieting and students' awareness and expectations about the environmental issues of deforestation and species extinction in the Middle East and North America: An integrated cross cultural ecologic-economic analysis" [5]. The present article presents key information and indicators on the world youth population before and after United Nations (UN) country members' adoption of the 2030 Agenda for sustainable development in 2015. In doing so, it provides about a five year window of facts and figures, which can be used by researchers and policy makers to evaluate the effectiveness of early stage implementation of SDGs strategies in relation to youth education, access to information and communication technologies (ICT), and well-being worldwide. The presented data covers 409747 youth respondents distributed across 46 countries, 199511 of the respondents are from the 2015 cross-section, and the remaining 210236 from the 2018 cross-section. The data is further supplemented with spatial metadata containing the geographical coordinates of each of the covered 46 countries, which can be used for spatial Analysis and econometric modeling as illustrated in the present data article.

Keywords

ICT resources; Middle East; OECD; PISA data; Spatial Analysis; Sustainable Development Goals; Youth Well-being

Specifications Table

Subject	Economics and Econometrics
Specific subject area	Development and Welfare Economics
Type of data	Table
How data were acquired	The main data files- from the SAS versions of the publicly released student questionnaire data files of the 2015 [2] and 2018 [3] Program for International Student assessment (PISA); The country level geospatial data-from the GADM database of Global Administrative Areas [6]. All data extractions were acheived using the R statistical software [9].
Data format	Raw Filtered R formatted Data frames and lists
Parameters for data collection	The criteria for variables selection was defined as the set of all ICT related repeated youth measurements in both the 2015 and 2018 cycles of the PISA, followed by the exclusion of all youth respondents from countries not represented in both cycles.
Description of data collection	The main R data object "Pool_Panel58" presented in this article was specifically extracted from the raw SAS data files "cy6_ms_cmb_stu_qqq.sas7bdat" and "cy07_msu_stu_qqq.sas7bdat" in the publicly released folders of the 2015 and 2018 "student questionnaire data files" respectively. The country level spatial metadata were collected from the GADM database of Global Administrative Areas [6].
Data source location	Institution: The Organization for Economic Cooperation and Development (OECD) Country: 46 (OECD and Non-OECD) countries worldwide as listed in Table (1) and further mapped in figure (1) below.
Data accessibility	With the article And also available with: Repository name: Mendeley Data [11] Direct URL to data: *http://dx.doi.org/10.17632/pntxmgf8td.1*

Related research article	Niankara, I. (2019). “Cross-national Data Sample on the Environmental Affection and Cognition of Adolescent Students of Varying Interests in Ecosystem Services and Sustainability”. Data in Brief, Vol. 22, February 2019, pp. 312-318. [1] Niankara, I. (2019). “Scientific media dieting and youth awareness and expectations about the environmental issues of deforestation and species extinction in the middle east and north America” World Reviews of Science, Technology and Sustainable Development. Vol. 15, No 3, pp. 252-282. [5]

Value of the Data

- This data allows researchers to examine within nations, as well as between nations changes in youth outcome measures, including Access to ICT resources, subjective well-being, and educational performance
- Due to its timing, it also provides a unique opportunity for evaluating the effectiveness of youth related SDG strategies that have been in implementation since 2015, under Goal 9, target 9C; Goal 4, target 4.4; and Goal 8, target 8.6 among others.
- The data also includes spatial Meta data, in the form of country map coordinates, for all 46 covered countries, which allows prospective investigators to address research questions within both, the time dimension and the spatial dimension.
- Specifically for example, using the Difference-in-difference estimation, one can rely on this data to uncover Differences in Youth Access to ICT resources and/or Subjective well-being in the SDG era across the 46 Nations and/or Economic Blocks (OECD countries vs Non-OECD countries).

Data Description

This data article is being provided to update and extend the previously published data article [1]. The article updates and extends the former because it includes data extract of the most recently released 2018 “Programme for International Student Assessment” (PISA) [3], in addition to data extract from the 2015 PISA [2]. In fact, the 2015 and 2018 cross-sections are pulled into a Panel data, which allows for investigations in more dimensions than the former published cross-sectional data of the 2015 PISA. Another key aspect on which the current data article extends the former article [1], is in its provision of country level geospatial metadata that are useful for spatial econometric modelling, and geographical mapping of country level youth response data aggregates. Overall, the present data article provides six R data objects, which are:

- “Pool_Panel58” that contains the pooled panel of youth respondents’ information, which are described in more details in table (2) and table (3) below.
- “Xsec_2015” and “Xsec_2018” represent respectively the 2015 and 2018 cross- sections of the recorded youth respondents’ information. Except for the data structure, the contents in terms of variables are identical to the description provided in table (2) and Table (3).
- “Pool_Panel58_tmap3” contains youth respondents information aggregated across both “country” and “year”, and further augmented with “.sf” spatial Meta data (shape files) from the GADM library [6]. These shape files are lists of spatial coordinates of the 46 countries, and are useful for the geographical mapping of the aggregated youth response data. This version of the aggregated data is useful for spatial analysis of country level youth responses over the full panel coverage (2015 to 2018), without cross-sectional distinction.
- “Pool_Panel58_tmap2” contains youth respondents information aggregated across “country” only, and augmented with country spatial coordinates for geographical mapping. This version of the aggregated data is useful for spatial analysis of country level youth responses in 2015 and 2018 distinctively.
- “Pool_Pan.plys” contains a list of 46 shape polynomial files, representing each of the 46 countries in the provided sample of youth respondents’ information “Pool_Panel58”. It is derived from the raw individual country “.sp” files in the GADM library [6]. This polynomial data file is required if one wishes to implement a spatial econometric analysis using the shared data “Pool_Panel58”, as is the case in [7].

Recall that the initial data file “Pool_Panel58” is a pooled panel of the student-questionnaire data files from 2015 and 2018 PISA, covering youth respondents’ information from 46 countries worldwide. As a triennial survey of adolescent students around the world, PISA was lunched by the Organization for Economic Co-operation and Development (OECD), to assess the degree of preparedness of students near the end of their compulsory education, for full participation in modern societies.

The geographical coverage of the shared data is summarized in figure (1), which maps the global count of youth respondents in the pooled sample. For further clarity in the contribution of each country to the final sample, figure (2) and figure (3) map the annual cross-country counts of youth respondents for the years 2015 and 2018 respectively. Supplementing the geographical mappings of the data, are the country level data summaries in Table (1), which shows the absolute frequency and percent relative frequency distributions of youth respondents across the 46 countries in the data. The first column summarizes the 2015 sub-sample, while the second column covers the 2018 sub-sample, and the third the pooled data sample.

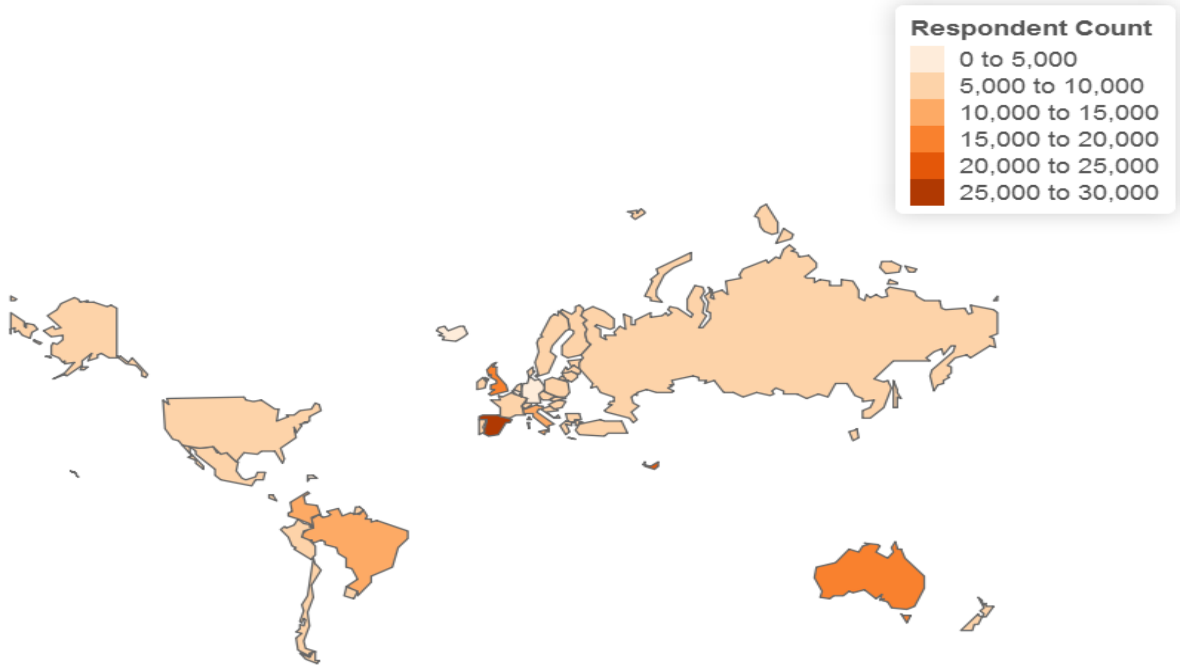

Figure 1: Global mapping of respondents count by country over the two years 2015 and 2018 (See dynamic web link at http://rpubs.com/brassbe1982/globyouthcount_fig1)

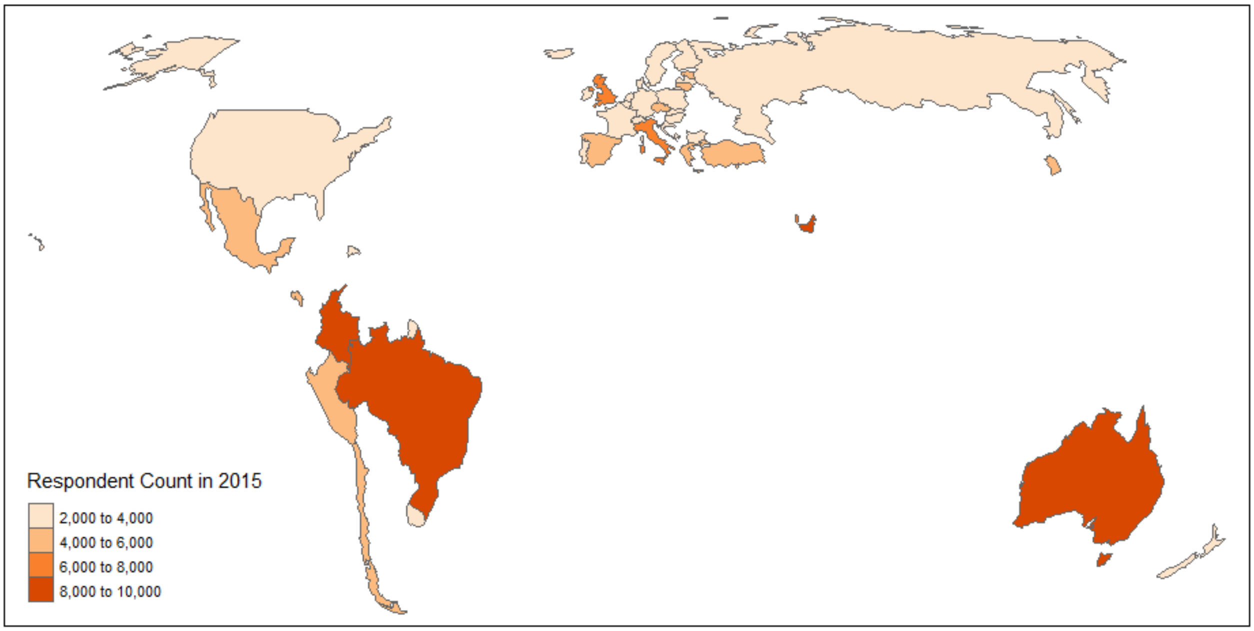

Figure 2: Global mapping of youth respondents count by country for the year 2015 (See dynamic web link at http://rpubs.com/brassbe1982/glob2015count_fig2)

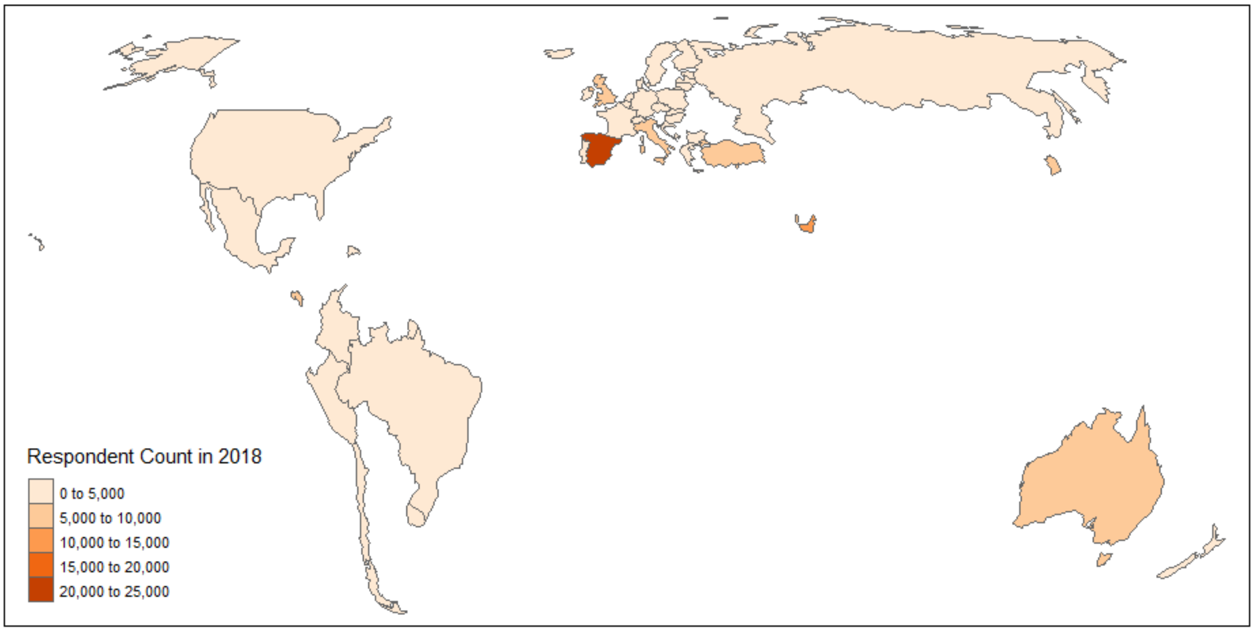

Figure 3: Global mapping of youth respondents count by country for the year 2018 (See dynamic web link at http://rpubs.com/brassbe1982/glob2018count_fig3)

Table (1) frequency and percent relative frequency distributions of youth respondents

N	**2015 Sub-sample**		**2018 Sub-sample**		**Combined 2015 + 2018**	
409747	Abs. Freq.	Rel. Freq.	Abs. Freq.	Rel. Freq.	Abs. Freq.	Rel. Freq.
Australia	8251	4.14%	7938	3.78%	16189	3.95%
Belgium	2388	1.20%	2144	1.02%	4532	1.11%
Brazil	9528	4.78%	4618	2.20%	14146	3.45%
Bulgaria	3351	1.68%	2453	1.17%	5804	1.42%
Chile	4296	2.15%	3866	1.84%	8162	1.99%
Chinese Taipei	4853	2.43%	5107	2.43%	9960	2.43%
Columbia	8554	4.29%	4766	2.27%	13320	3.25%
Costa Rica	4279	2.14%	5100	2.43%	9379	2.29%
Croatia	3756	1.88%	4316	2.05%	8072	1.97%
Czech Republic	4306	2.16%	4029	1.92%	8335	2.03%
Denmark	2716	1.37%	3848	1.83%	6586	1.61%
Dominican Republic	2555	1.28%	1168	0.56%	3723	0.91%
Estonia	4164	2.09%	3546	1.69%	7710	1.88%
Finland	3882	1.95%	3404	1.62%	7286	1.78%
France	3765	1.89%	3495	1.66%	7260	1.77%
Germany	2616	1.31%	1719	0.82%	4335	1.06%
Greece	4150	2.08%	4510	2.15%	8660	2.11%
Hong Kong	2676	1.34%	4191	1.99%	6867	1.68%
Hungary	3170	1.59%	3338	1.59%	6508	1.59%
Iceland	2203	1.10%	1855	0.88%	4058	0.99%
Ireland	3839	1.92%	3468	1.65%	7307	1.78%
Italy	7866	3.94%	6565	3.12%	14431	3.52%
Korea	4450	2.23%	5454	2.59%	9904	2.42%
Latvia	3172	1.59%	3144	1.50%	6316	1.54%

Table (1) frequency and percent relative frequency distributions of youth respondents (Continue)

N	**2015 Sub-sample**		**2018 Sub-sample**		**Combined** **2015 + 2018**	
409747	Abs. Freq.	Rel. Freq.	Abs. Freq.	Rel. Freq.	Abs. Freq.	Rel. Freq.
Lithuania	4102	2.06%	3834	1.82%	7936	1.94%
Luxembourg	3104	1.56%	3250	1.55%	6354	1.55%
Macao	3286	1.65%	3199	1.52%	6485	1.58%
Mexico	5637	2.83%	3593	1.71%	9230	2.25%
Montenegro	3143	1.58%	4317	2.05%	7460	1.82%
Netherlands	3084	1.55%	2370	1.13%	5454	1.33%
New Zealand	2420	1.21%	3511	1.67%	5931	1.45%
Peru	5059	2.54%	2293	1.09%	7352	1.79%
Poland	3431	1.72%	4052	1.93%	7483	1.83%
Portugal	3429	1.72%	3294	1.57%	6723	1.64%
Qatar	6703	3.36%	9212	4.38%	15915	3.88%
Russian Federation	3976	1.99%	4873	2.32%	8849	2.16%
Singapore	4485	2.25%	4803	2.28%	9288	2.27%
Slovak Republic	3904	1.96%	3396	1.62%	7300	1.78%
Spain	4240	2.13%	21518	10.24%	25758	6.29%
Sweden	3440	1.72%	3387	1.61%	6827	1.67%
Switzerland	3317	1.66%	2808	1.34%	6125	1.49%
United Arab Emirates	8937	4.48%	13409	6.38%	22346	5.45%
Turkey	4123	2.07%	5819	2.77%	9942	2.43%
United Kingdom	7910	3.96%	7718	3.67%	15628	3.81%
United States	3822	1.92%	3272	1.56%	7094	1.73%
Uruguay	3151	1.58%	2266	1.08%	5417	1.32%

Experimental Design, Materials, and Methods

The main data file “Pool_Panel58” that contain youth level response information was extracted from the raw SAS (TM) version of the “student questionnaire data files” of the 2015 and most recent 2018 PISA. These raw files are released, as zipped folders under the names “PUF_SAS_COMBINED_CMB_STU_QQQ.zip” [2] and SAS_STU_QQQ.zip [3].

These zipped folders are very large data folders each with raw SAS formatted data files “.sas7bdat”, along with their respective description files “.sas7bdat.format.sas”.

The R data object “Pool_Panel58” was specifically extracted from the SAS data files “cy6_ms_cmb_stu_qqq.sas7bdat” and “cy07_msu_stu_qqq.sas7bdat” in the shared folders of the 2015 and 2018 “student questionnaire data files” respectively. Due to the very large size of the raw data files, and the memory constraint imposed by the R statistical software, the attached R extraction codes were used to import only the subsets of each raw data file containing the variables of primary interest. The criteria for variables selection was defined as the set of all ICT (Information and Communication Technology) related repeated youth measurements in both, the 2015 and 2018 cycles of the PISA.

These measurements were pooled into a single R data object “data58N”. Since we wanted a version of the queried data that would not only allow for inter-temporal, but also inter-and-intra-country youth level outcomes analyses, we dropped from the final pooled sample “Pool_Panel58”, all youth respondents from countries not represented in both cycles of the PISA. This last version of the pooled data “Pool_Panel58”, obtained after subsequent treatments of “data58N”, covers 409747 youth respondents distributed across 46 countries worldwide. The 2018 extract of this last version covering a set of five countries in the Middle East region was used for the analysis in [7]. This data article presents however, the full data set. For further details on the sampling design of the PISA, please consult the OECD report [8].

Below we describe the experimental factors contained in the pooled data “Pool_Panel58”. Table (2) shows the descriptive statistics (mean and standard deviations) for select key quantitative measurements, while table (3) presents the descriptive statistics (absolute frequency and percent relative frequency) for select key qualitative measurements. All data treatments, variable reformulation, and descriptive statistics were implemented within the R statistical Software [9]; (see the attached R computer codes, for more details).

In addition to the description of the variables in table (2) and table (3), using the R library “dplyr” [10], we provide descriptive analytics of country level aggregated youth outcomes. The result of these treatment are summarized in the R data objects “Pool_Panel58_tmap2” and “Pool_Panel58_tmap3” as previously described in the data section above. For more illustration, Figures (4-9) below provide the geographical mapping of a select number of key country level aggregated youth responses from “Pool_Panel58_tmap3”. Further details on the initial non-aggregated indices and their interpretations can also be found in section 3.1.1 of [7].

Table 2: Descriptive statistics for select key quantitative measurements in “Pool_Panel58” (N =409747)

Quantitative variables	**Description**	**Mean**	**s.d.**
BELONG	Sense of belonging in school (index of youth subjective well-being)	-0.011	1.008
ICTRES	Index of ICT resources availability to youth at home	-0.125	1.071
nPhonInternAcH	Number of Phones with internet access at home	3.693	0.694
nCompH	Number of computers at home (Desktop, Laptop, notebook)	2.975	0.956
AGE	student's age in years	15.789	0.291
ExpecOccup	student's expected occupational status by age 30	64.698	18.306
HEDRES	Index of home educational resources availability	-0.064	1.025
CULTPOSS	Index of cultural possessions at home	-0.047	0.973
HISCED	Index of highest education of parents (international standard classification of education - ISCED)	4.754	1.445
ESCS	Standardized PISA Index of economic, social and cultural status.	-0.073	1.028
HISEI	Index of highest parental occupational status	52.674	22.085
W_FSTUWT	Student final weight in the Data	48.241	105.470

Note : s.d. represents the standard deviation of the quantitative measurement

Table 3 : Descriptive statistics for select key qualitative measurements in “Pool_Panel58” (N =409747)

Qualitative variables	**Levels**	**Abs. Freq.**	**Rel. Freq.**
InternetLink	0-No link to the internet at home	24508	5.98%
	1-Yes has link to the internet at home	385239	94.02%
Gender	0- Male	195688	47.76%
	1-Female	214059	52.24%
IMMIG	Respondent’s immigration status		
	1-Native	348269	85.00%
	2-Second-generation	29988	7.32%
	3- First-generation	31490	7.69%
GradeLev	Student grade level in School		
	7th grade	2015	0.49%
	8th grade	14664	3.58%
	9th grade	135224	33.00%
	10th grade	208150	80.80%
	11th grade	47084	11.49%
	12th and 13th grade	2610	0.64%
LangH	Most spoken language at home		
	0-other languages	52377	12.78%
	1- language of the test	357370	87.22%

Note: Abs. Freq. and Rel. Freq. represent respectively the absolute frequency and relative frequency distributions of the qualitative variables

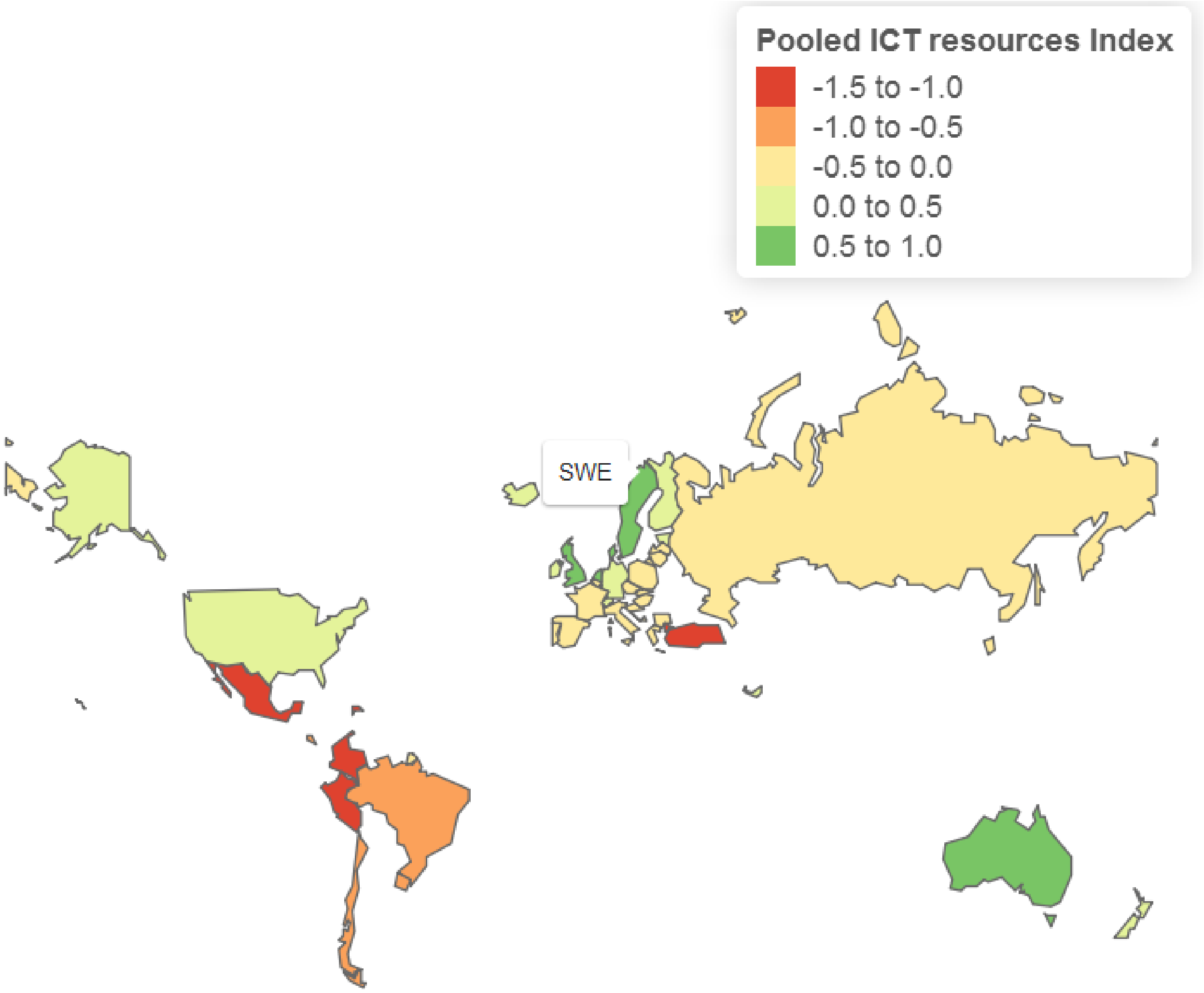

Figure 4: Spatial distribution of pooled (2015 and 2018) country level weighted average of "Youth access to ICT resources".
(See dynamic web link at http://rpubs.com/brassbe1982/ICTRESindex_fig4)

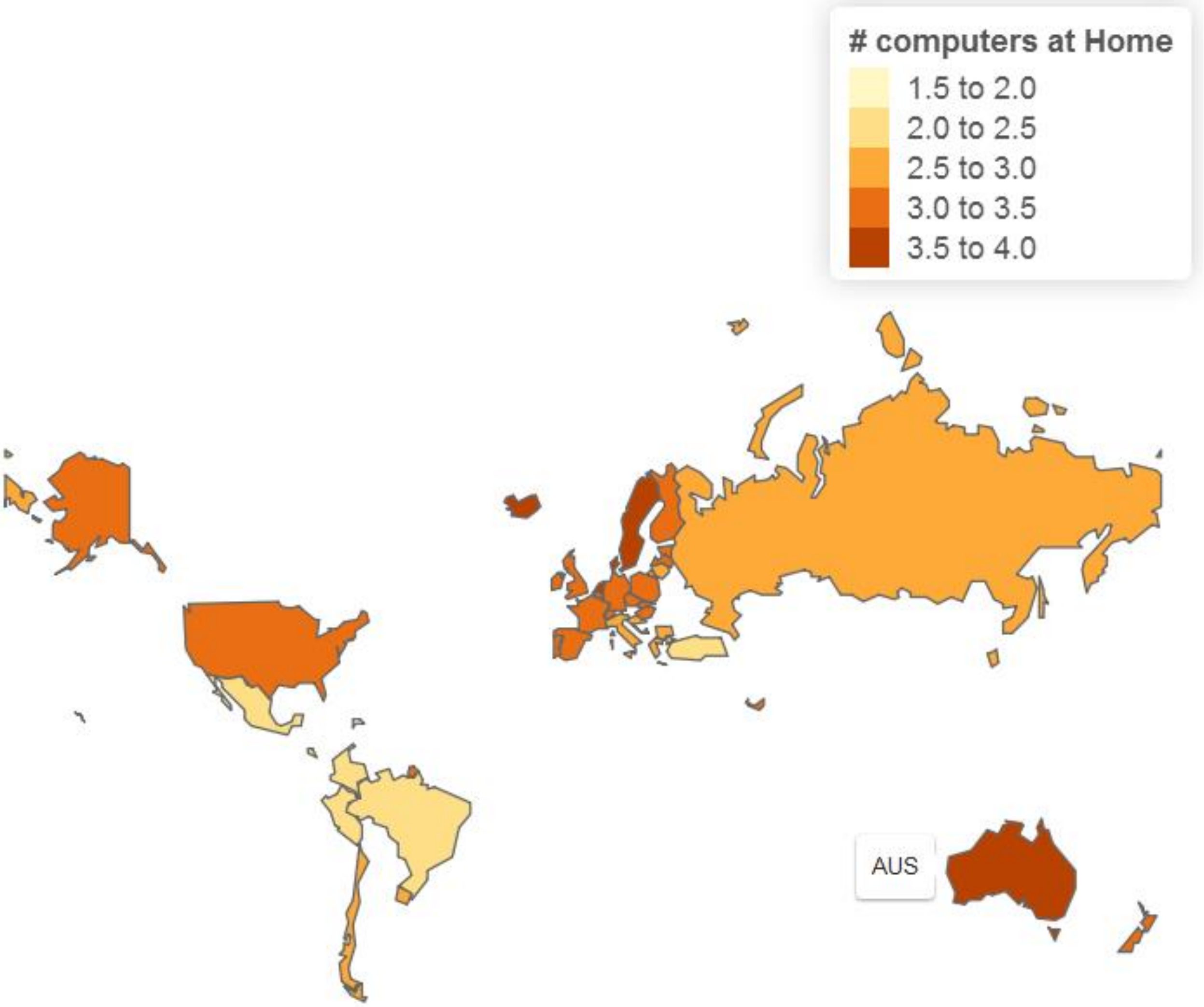

Figure 5: Spatial distribution of pooled (2015 and 2018) country level weighted average of "number of computers at youth's home (desktop, laptop, and notebook)".
(See dynamic web link at http://rpubs.com/brassbe1982/nCompH_fig5)

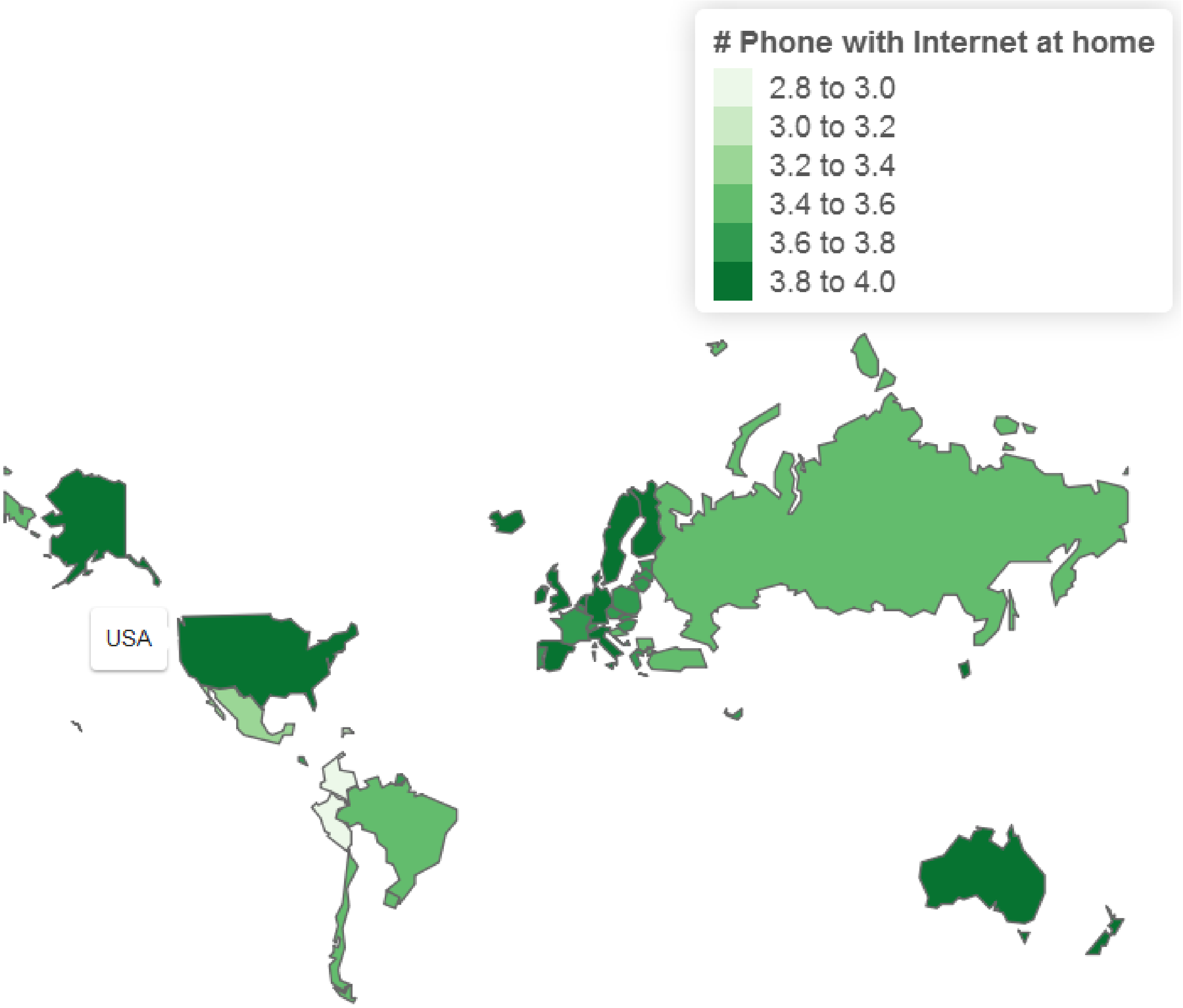

Figure 6: Spatial distribution of pooled (2015 and 2018) country level weighted average of "number of phones with internet access at youth's home".
(See dynamic web link at http://rpubs.com/brassbe1982/nPhonInternAcH_fig6)

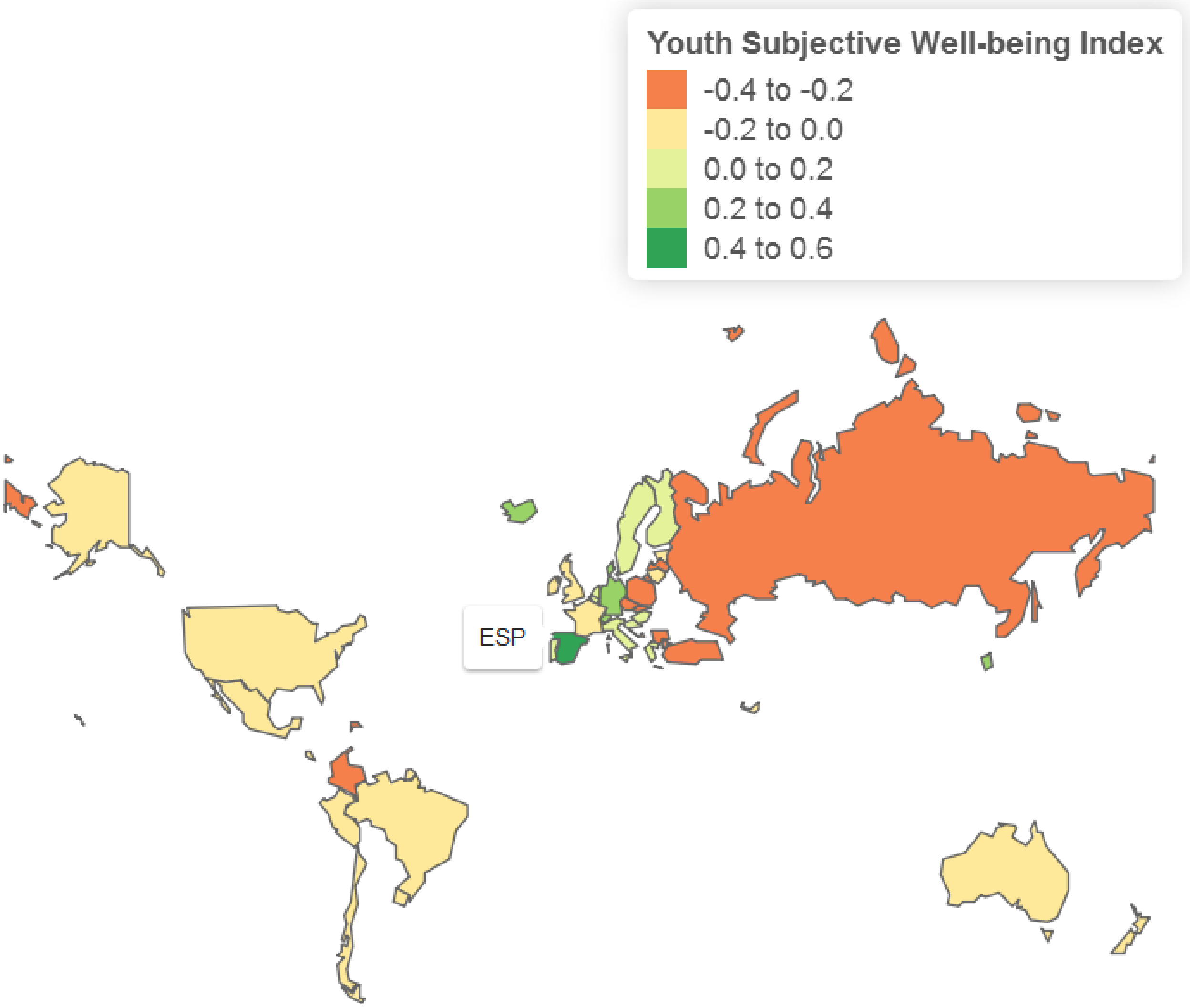

Figure 7: Spatial distribution of pooled (2015 and 2018) country level weighted average of "youth's sense of belonging in school".
(See dynamic web link at http://rpubs.com/brassbe1982/Belong_fig7)

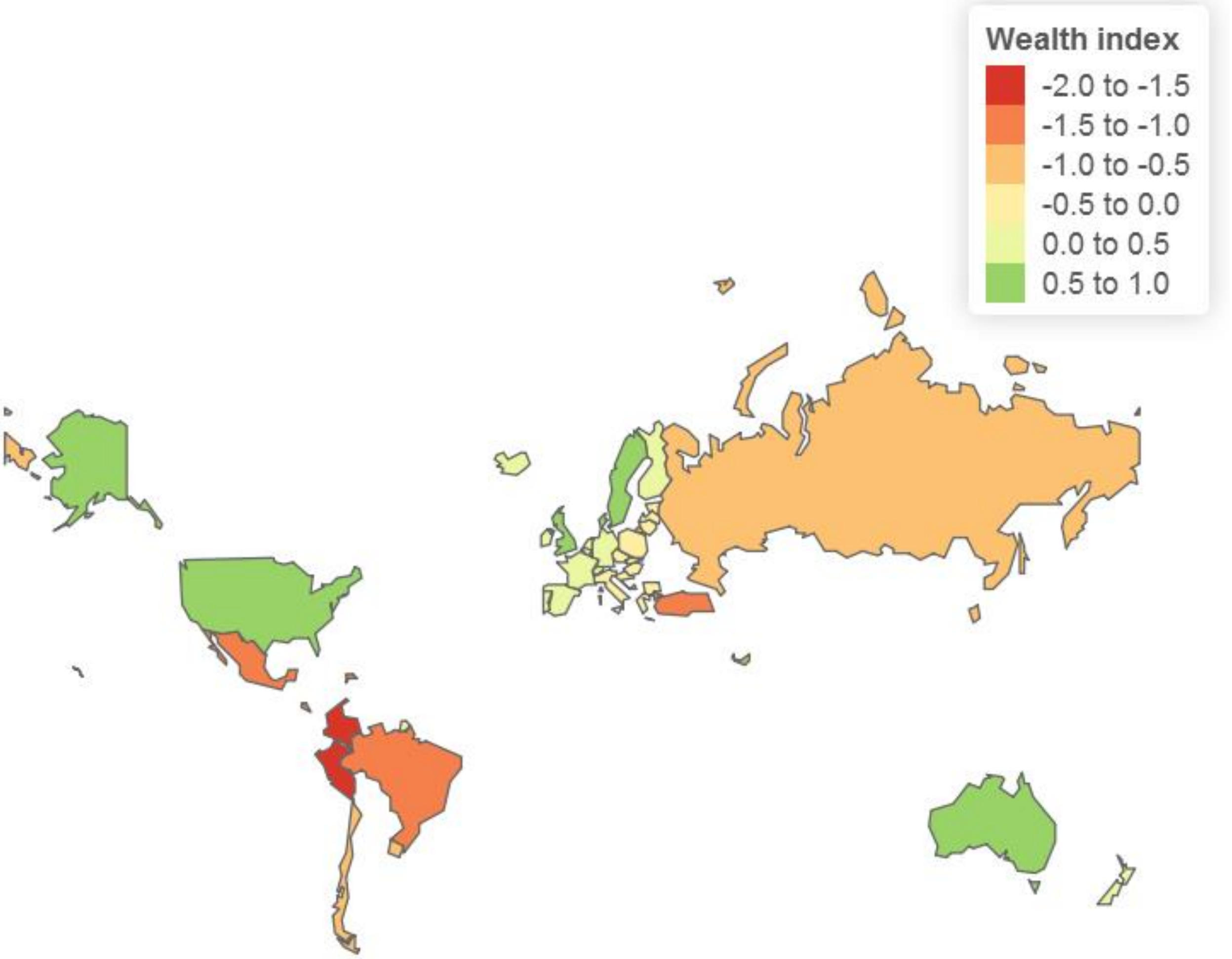

Figure 8: Spatial distribution of pooled (2015 and 2018) country level weighted average of "youth's family wealth index".
(See dynamic web link at http://rpubs.com/brassbe1982/WealthIndex_fig8)

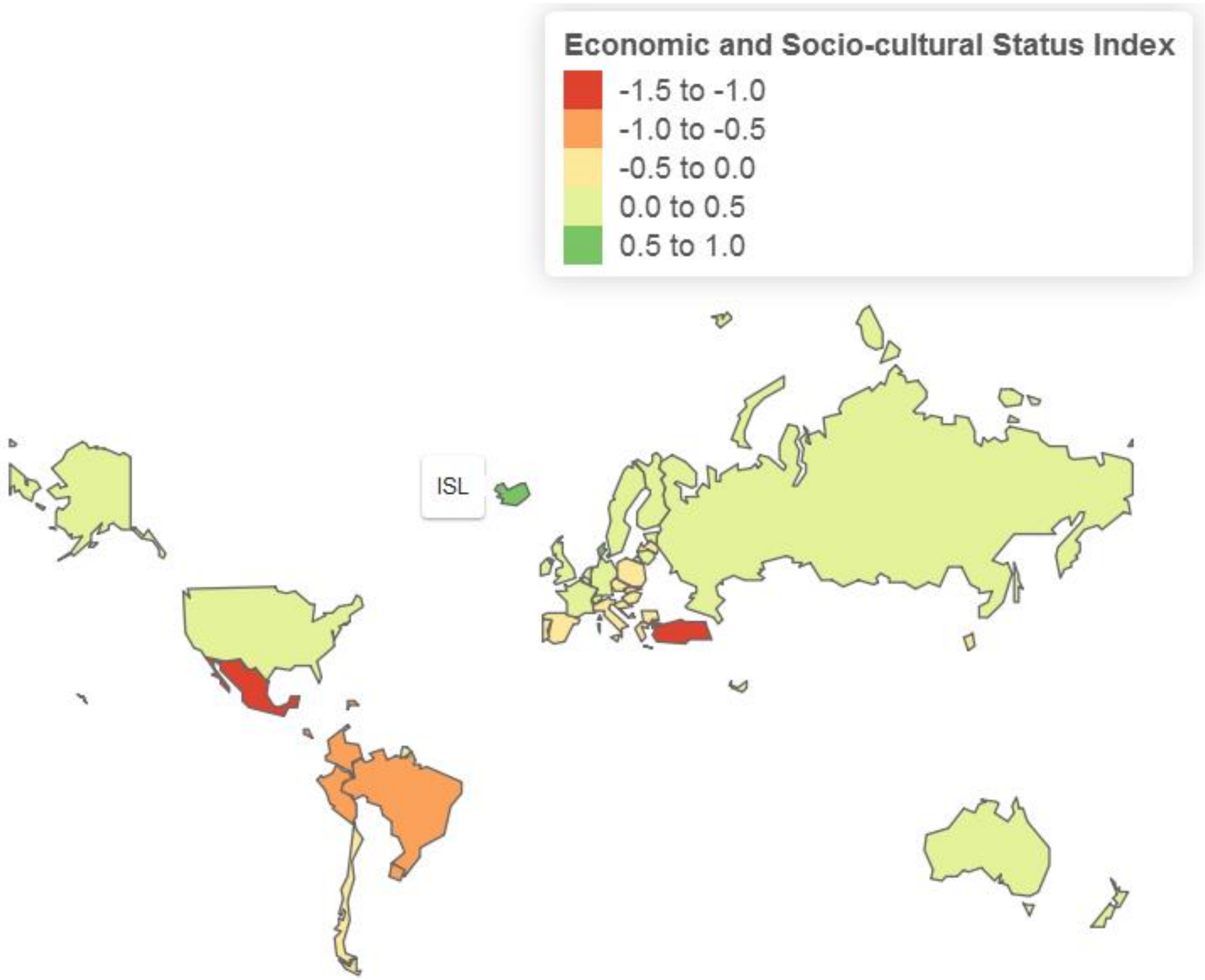

Figure 9: Spatial distribution of country level weighted average of "youth's index of Economic and Socio-cultural Status".
(See dynamic web link at http://rpubs.com/brassbe1982/ESCSindex_fig9)

Acknowledgments

The Authors of this research work express their sincere appreciation to the Organization for Economic Co-operation and Development (OECD) for the initial collection and sharing of the raw public use data files of the Programme for International Student Assessment (PISA).

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that have, or could be perceived to have, influenced the work reported in this article.

References

[1] Niankara, I. (2019). “Cross-national Data Sample on the Environmental Affection and Cognition of Adolescent Students of Varying Interests in Ecosystem Services and Sustainability”. *Data in Brief, Vol. 22, February 2019, pp. 312-318.*

[2] OECD (2016), Programme for International Student Assessment (PISA) 2015 Database: Student Questionnaire data file, Organization for Economic Co-operation and Development, Paris, France. Retrieved on 30 March 2018 from http://www.oecd.org/pisa/data/2015database/

[3] OECD (2019), Programme for International Student Assessment (PISA) 2018 Database: Student Questionnaire data file, Organization for Economic Co-operation and Development, Paris, France. *Retrieved on 10 December 2019 from* http://www.oecd.org/pisa/data/2018database/

[4] Niankara I. and Zoungrana, D. T. (2018) “Interest in the Biosphere and students environmental awareness and optimism: A global perspective”, *Global Ecology and Conservation, Vol. 16, e00489.*

[5] Niankara, I. (2019). “Scientific media dieting and youth awareness and expectations about the environmental issues of deforestation and species extinction in the middle east and north Ameri*ca” World Reviews of Science, Technology and Sustainable Development. Vol. 15, No 3, pp. 252-282.*

[6] Hijmans, R., Garcia, N., & Wieczorek, J. “GADM database of Global Administrative Areas”. *Version 3.6 (released May 6, 2018). [Online] Retrieved March 12, 2019 from* https://gadm.org/data.html

[7] Niankara, I. (2020). “Cross national comparative analysis of youth access to ICT resources and subjective well-being in the Middle East: A Spatial bivariate copula regression modelling” *in preparation for the 2020 world congress of the econometrics society. Milan, Italy. (August 17-21). (see attached supplementary materials).*

[8] OECD (2017), PISA 2015 Technical Report: Chapter 04 - Sample design, Organization for Economic Co-operation and Development, Paris, France. *Retrieved on 30 March 2018 from* http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

[9] R Core Team, “R: A Language and Environment for Statistical Computing”, *R Foundation for Statistical Computing, Vienna, Austria,* 2015. *URL:* https://www.R-project.org/

[10] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. *R package version 0.8.0.1. [Online] Retrieved March 12, 2019 from* https://CRAN.R-project.org/package=dplyr

[11] [Dataset] Niankara, Ibrahim (2020), “Data for cross national comparative analysis of youth access to ICT resources and subjective well-being in the Middle East: A spatial bivariate copula regression modelling”, *Mendeley Data, v1 http://dx.doi.org/10.17632/pntxmgf8td.1*