The Journal of Open Source Software

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

spiR, An R package to access social progress data

Thierry Warin[1, 2]

1 Professor, HEC Montral (CANADA) **2** Principal Investigator, CIRANO (Montreal, Canada)

Summary

The spiR package is intended to make two simple contributions: (1) to give access to new societal and economic data and (2) to integrate easily in a researcher's workflow through the R language.

By providing new open data in an integrated way, it highlights some other principles: open data, open code, open and reproducible science. The novelty of the data, collected and computed by the Social Progress Imperative whose advisory board is chaired by Professor Michael Porter (Harvard Business School), helps inform better policy decisions when it comes to accelerate efforts to drive equitable, inclusive and thriving societies.

Updated on a quarterly basis, the R package contains 72 metadata, covers 194 countries from 2014 and is rich of 1434 observations. It is also expanding by providing access to other data sources in order to augment the level of geographic granularity, while making sure data comparisons are consistent.

Motivation

The spiR package was developped to facilitate the access to new indicators measuring social progress while being in the open data spirit. 'spiR' is an R package to easily access the Social Progress Index datasets. It is inspired by several other initiatives in the open data space (Kim et al., 2019).

This R package provides a collection of new indicators to compute complexity metrics that offer an easy access to these data for social scientists (Vargas, 2020). Indeed, in the past decade, a new field has emerged: Economic Complexity. Economic Complexity is a data-driven approach that may be used to inform the territorial development policies with evidence-based metrics. An interdisciplinary analytical framework by essence, Economic complexity stands at the crossroads between evolutionary economics and institutional economics (Cimoli & Dosi, 1995; Hirschman, 1958; Teece, Rumelt, Dosi, & Winter, 1994). It was originally designed to analyze the determinants of economic development. Driven by an inferential approach, the need for new indicators and new data has occurred. In our current technological "state of the art", it is important for social scientists to take advantage of the new possibilities. Social scientists can leverage technologies to have access to new kinds of data (structured and unstructured) and compute new indicators. The goal is the same: to better understand - and hence inform - individuals in their interactions with societies and societies themselves.

The differences with the previous literature are essentially twofold. First, at the theoretical level, this literature proposes to consider different dimensions (industries and product spaces) shifting the focus away from only aggregate variables (Hidalgo & Hausmann, 2009; Tacchella, Cristelli, Caldarelli, Gabrielli, & Pietronero, 2012). The second difference is the result of researchers having a better access to computing power and data. Indeed, with respect to the previous literature, Economic Complexity is by nature data-intensive and compting-power intensive. With the current acceleration in computing power access and open data initiatives across the world, Economic Complexity can now be used to provide complementary analyses to more conventional macroeconomic methods. The shift in granularity allowed by this new access to data allows researchers to answer quantitatively to several policy relevant questions.

In the field of Economic Complexity, open data initiatives constitute essential resources. Across the world, the question of facilitating the access to socioeconomic data is widespread and well established nowadays. While people are more and more connected, the need to provide reliable sources of information as well as providing validated data to form better decisions has become crucial and well understood. In this context, governments, research centers, private companies, NGOs and think tanks at different levels have started to build open data initiatives. Another related goal is to rely on reproducible research principles. By building up a workflow of integrated tools such as data, code and methods, a researcher can share her results more widely in a reproducible spirit.

In 2015, The 17 United Nations' Sustainable Development Goals were adopted. 'spiR' is a package comprised of several open datasets, which are published by the Social Progress Imperative (https://www.socialprogress.org/), including the Social Progress Index (a synthetic measure of human development across the world) (Progress, 2020). 'spiR''s goal is to provide data to help policymakers and researchers prioritize actions that accelerate social progress across the world in the context of the Sustainable Development Goals. The Social Progress Index proposes a new perspective on social challenges and needed efforts to accelerate social progress in line with the Sustainable Development Goals. In this context, the goal of 'spiR' is to allow an easy connection with R to the Social Progress Index.

At the Social Progress Imperative, they define "social progress as the capacity of a society to meet the basic human needs of its citizens, establish the building blocks that allow citizens and communities to enhance and sustain the quality of their lives, and create the conditions for all individuals to reach their full potential. Improving quality of life is a complex task and past efforts to measure progress simply haven't created a sufficiently nuanced picture of what a successful society looks like. That's why we created the Social Progress Index. Rather than emphasizing traditional measurements of success like income and investment, we measure 51 social and environmental indicators to create a clearer picture of what life is really like for everyday people. The index doesn't measure people's happiness or life satisfaction, focusing instead on actual life outcomes in areas from shelter and nutrition to rights and education. This exclusive focus on measurable outcomes makes the index a useful policy tool that tracks changes in society over time."

Having an easy access to these data has an outstanding impact on education and research. New research questions can be answered thanks to the new level of data availabilitym while being able to tap into the "wisdom of crowds" (Gal (2019), Cai, Gippel, Zhu, & Singh (2019), De C. Wang et al. (2019), Avasilcai & Galateanu (2018)). In the past, there were lots of areas in public policy-making where data were not accessible. As a result, decisions were made on asssumptions coming from theoretical foundations or from benchmarks from other sources. Numerous authors have demonstrated the role of data in informing better evidence-based policies (Wolffe, Whaley, Halsall, Rooney, & Walker (2019), Payán & Lewis (2019), Giménez-bertomeu, Domenech-lópez, Mateo-pérez, & De-alfonseti-hartmann (2019), Villumsen, Faxvaag, & Nøhr (2019)).

Functionality and Reproducible demonstration

The index measures the quality of life for 98% of the world's population. In its current version, the R package provides access to global data. In further versions, this package will include different geographical levels: states, regions, cities and sometimes communities.

We will also include education and innovation indicators to capture the dynamics of social progress across the world (Hadengue & Warin, 2014).

Three overarching dimensions are (1) Basic Human Needs, (2) Foundations of Wellbeing, and (3) Opportunity. Within each dimension, there are four components that further divide the indicators into thematic categories. The index consists in 51 social and environmental variables, covering the years 2014 to 2019.

spiR covers 72 metadata (mostly economic and societal variables) and contain 1434 observations for each metadata from 2014 to 2019 across 194 countries. These data are easily accessible through three functions: (1) to select the country, (2) to select the indicator(s) and (3) to collect the data in a directly usable dataframe.

Indicator	Code
Social Progress Index	SPI
Basic Human Needs	BHN
Foundations of Wellbeing	FOW
Opportunity	OPP
Nutrition and Basic Medical Care	NBM
Water and Sanitation	WSA
Shelter	SHE
Personal Safety	PSA
Access to Basic Knowledge	ABK
Access to Information and Communications	AIC
Health and Wellness	HWE
Environmental Quality	EQU
Personal Rights	PRI
Personal Freedom and Choice	PFC
Inclusiveness	INC
Access to Advanced Education	AAE
Undernourishment (% of pop.)	NBM_1
Maternal mortality rate (deaths/100,000 live births)	NBM_2
Child mortality rate (deaths/1,000 live births)	NBM_3
Child stunting (% of children)	NBM_4
Deaths from infectious diseases (deaths/100,000 people)	NBM_5
Access to at least basic drinking water (% of pop.)	WSA_1
Access to piped water (% of pop.)	WSA_2
Access to at least basic sanitation facilities (% of pop.)	WSA_3
Rural open defecation (% of pop.)	WSA_4
Access to electricity (% of pop.)	SHE_1
Quality of electricity supply (1=low; 7=high)	SHE_2
Household air pollution attributable deaths (deaths/100,000 people)	SHE_3
Access to clean fuels and technology for cooking (% of pop.)	SHE_4
Homicide rate (deaths/100,000 people)	PSA_1
Perceived criminality (1=low; 5=high)	PSA_2
Political killings and torture (0=low freedom; 1=high freedom)	PSA_3
Traffic deaths (deaths/100,000 people)	PSA_4
Adult literacy rate (% of pop. aged 15+)	ABK_1
Primary school enrollment (% of children)	ABK_2
Secondary school enrollment (% of children)	ABK_3
Gender parity in secondary enrollment (distance from parity)	ABK_4
Access to quality education (0=unequal; 4=equal)	ABK_5
Mobile telephone subscriptions (subscriptions/100 people)	AIC_1
Internet users (% of pop.)	AIC_2
Access to online governance (0=low; 1=high)	AIC_3

Indicator	Code
Media censorship (0=frequent; 4=rare)	AIC_4
Life expectancy at 60 (years)	HWE_1
Premature deaths from non-communicable diseases (deaths/100,000 people)	HWE_2
Access to essential services(0=none; 100=full coverage)	HWE_3
Access to quality healthcare (0=unequal; 4=equal)	HWE_4
Outdoor air pollution attributable deaths (deaths/100,000 people)	EQU_1
Greenhouse gas emissions (CO2 equivalents/GDP)	EQU_2
Biome protection	EQU_3
Political rights (0=no rights; 40=full rights)	PRI_1
Freedom of expression (0=no freedom; 1=full freedom)	PRI_2
Freedom of religion (0=no freedom; 4=full freedom)	PRI_3
Access to justice (0=non-existent; 1=observed)	PRI_4
Property rights for women (0=no rights; 5=full rights)	PRI_5
Vulnerable employment (% of employees)	PFC_1
Early marriage (% of women)	PFC_2
Satisfied demand for contraception (% of women)	PFC_3
Corruption (0=high; 100=low)	PFC_4
Acceptance of gays and lesbians (0=low; 100=high)	INC_1
Discrimination and violence against minorities (1=low; 10=high)	INC_2
Equality of political power by gender (0=unequal power; 4=equal power)	INC_3
Equality of political power by socioeconomic position (0=unequal power; 4=equal power)	INC_4
Equality of political power by social group (0=unequal power; 4=equal power)	INC_5
Years of tertiary schooling	AAE_1
Women's average years in school	AAE_2
Globally ranked universities (points)	AAE_3
Percent of tertiary students enrolled in globally ranked universities	AAE_4

The functions in spiR are:

- spi_country()
- spi_indicator()
- spi_data()

In three easy steps, a researcher can integrate spiR's data into her workflow.

First, the user needs to enter the ISO code of a country. To have access to this code, the following function provides this information: spi_country(). This function provides a list of all the countries available in spiR. spi_country(country = "Canada") provides the ISO code of the country.

Second, the user needs to specify which indicator is of interest. Again, spi_indicator() generates a list of all indicators. Then, it is possible to specify which indicator needs to be targeted: for instance, spi_indicator(indicators = "mortality") generates a list of indicators with "mortality" in their titles.

Third, once the user knows the ISO code and the indicator's code, she can collect the data in a very easy way through this function:

spi_data(country = c("USA", "FRA"), year = c("2018", "2019"), indicators = "SPI") # It generates a data frame of the overall SPI indicator for the USA and France for the years 2018 and 2019.

Other ways are possible, following the R grammar, for instance: spi_data(country = c("USA", "FRA"), years = "2018",) generates a data frame of all the indicators for the USA and France for the year 2018. The GitHub vignette (<www.github.com/warint/spiR>).

Availability

spiR is an open source software made available under the MIT license. It can be installed through the CRAN repository using: install.packages('spiR') or through the GitHub repository for the development version using the remotes package: remotes::install_github('warint/spiR').

Acknowledgements

The author would like to thank the Center for Interuniversity Research and Analysis of Organizations (CIRANO, Montreal) for its support, as well as Thibault Senegas, Marine Leroi and Martin Paquette at SKEMA Global Lab in AI. The usual caveats apply.

References

Avasilcai, S., & Galateanu, E. (2018). *Co-creators in innovation ecosystems. Part II: Crowdsprings 'Crowd in action. 400*. https://doi.org/10.1088/1757-899X/400/6/062001

Cai, C., Gippel, J., Zhu, Y., & Singh, A. (2019). The power of crowds: Grand challenges in the Asia-Pacific region. *Australian Journal of Management*, *44*(4), 551–570. https://doi.org/10.1177/0312896219871979

Cimoli, M., & Dosi, G. (1995). Technological Paradigms, Patterns of Learning and Development: An Introductory Roadmap. *Journal of Evolutionary Economics*, *5*(3), 243–268. Retrieved from https://econpapers.repec.org/article/sprjoevec/v_3a5_3ay_3a1995_3ai_3a3_3ap_3a243-68.htm

De C. Wang, P., Soares, V., De Souza, J., Esteves, M., Schots, N., & Duarte, F. (2019). *A crowd science framework to support the construction of a gold standard corpus for plagiarism detection*. 440–445. https://doi.org/10.1109/CSCWD.2019.8791853

Gal, M. (2019). The Power of the Crowd in the Sharing Economy. *Law and Ethics of Human Rights*, *13*(1), 29–59. https://doi.org/10.1515/lehr-2019-0002

Giménez-bertomeu, V., Domenech-lópez, Y., Mateo-pérez, M., & De-alfonseti-hartmann, N. (2019). Empirical evidence for professional practice and public policies: An exploratory study on social exclusion in users of primary care social services in Spain. *International Journal of Environmental Research and Public Health*, *16*(23). https://doi.org/10.3390/ijerph16234600

Hadengue, M., & Warin, T. (2014). Patterns of Specialization and (Un)Conditional Convergence: The Cases of Brazil, China and India. *Management International / International Management / Gestiòn Internacional*, *18*, 123–141. https://doi.org/https://doi.org/10.7202/1027869ar

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575. https://doi.org/10.1073/pnas.0900943106

Hirschman, A. (1958). *The strategy of economic development*. Retrieved from https://books.google.ca/books?id=wls-AAAAYAAJ

Kim, H., D'Orazio, V., Brandt, P., Looper, J., Salam, S., Khan, L., & Shoemate, M. (2019). UTDEventData: An R package to access political event data. *Journal of Open Source Software*, *4*(36), 1322. https://doi.org/10.21105/joss.01322

Payán, D., & Lewis, L. (2019). Use of research evidence in state health policymaking: Menu labeling policy in California. *Preventive Medicine Reports*, *16*. https://doi.org/10.1016/j.pmedr.2019.101004

Progress, S. (2020). 2019 Social Progress Index. Retrieved April 29, 2020, from https://www.socialprogress.org/

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2012). A New Metrics for Countries' Fitness and Products' Complexity. *Scientific Reports*, *2*(1), 1–7. https://doi.org/10.1038/srep00723

Teece, D. J., Rumelt, R., Dosi, G., & Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, *23*(1), 1–30. https://doi.org/10.1016/0167-2681(94)90094-9

Vargas, M. (2020). Economiccomplexity: Computational Methods for Economic Complexity. *Journal of Open Source Software*, *5*(46), 1866. https://doi.org/10.21105/joss.01866

Villumsen, S., Faxvaag, A., & Nøhr, C. (2019). Development and progression in Danish eHealth policies: Towards evidence-based policy making. *Studies in Health Technology and Informatics*, *264*, 1075–1079. https://doi.org/10.3233/SHTI190390

Wolffe, T., Whaley, P., Halsall, C., Rooney, A., & Walker, V. (2019). Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environment International*, *130*. https://doi.org/10.1016/j.envint.2019.05.065