SoftwareX 21 (2023) 101297

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

hdData360r: A high-dimensional panel data compiler for governance, trade, and competitiveness indicators of World Bank Group platforms

Marcell T. Kurbucz *

Department of Computational Sciences, Wigner Research Centre for Physics, 29-33 Konkoly-Thege Miklós Street, Budapest, H-1121, Hungary
Institute of Data Analytics and Information Systems, Corvinus University of Budapest, 8 Fővám Square, Budapest, H-1093, Hungary

ARTICLE INFO

Article history:
Received 13 September 2022
Received in revised form 9 December 2022
Accepted 22 December 2022

Dataset link: https://data.mendeley.com/datasets/jwkk44trj3

Keywords:
Data compiler
Panel data
Spatial data
Governance
Trade
Competitiveness

ABSTRACT

The World Bank Group's GovData360 and TCdata360 platforms are widely employed in socio-economic research. The thousands of governance, trade, and competitiveness indicators they contain served as the basis for much research in the field of economic, developmental, and cultural studies, coronavirus disease 2019-related research, and tourism, to state a few. The presented R package called `hdData360r` collects thousands of up-to-date annual indicators from these platforms for all countries worldwide. Furthermore, it allows missing value imputation with data from previous years, and optionally, it exports the generated dataset into tab-separated value (`TSV`) files. The `hdData360r` R package with a sample dataset it generates is available publicly on GitHub and Mendeley Data.

Code metadata

Current code version	v0.1.0
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-22-00282
Code Ocean compute capsule	Not applicable
Legal Code License	GNU General Public License v3.0
Code versioning system used	Git
Software code languages, tools, and services used	R
Compilation and installation requirements, operating environments & dependencies	R 4.1.3 or later, dependencies: `lubridate`, `data360r`, `stringr`, `dplyr`, `zoo` (if they have not been previously installed, these packages are installed automatically when the function is called). OS agnostic (Linux, OS X, MS Windows).
Link to developer documentation and user manual	https://github.com/mtkurbucz/hdData360r/blob/main/README.md
Support email for questions	kurbucz.marcell@wigner.hu

Software metadata

Current code version	v0.1.0
Permanent link to code/repository used for this code version	https://github.com/mtkurbucz/hdData360r
Legal Code License	GNU General Public License v3.0
Compilation and installation requirements, operating environments & dependencies	R 4.1.3 or later, dependencies: `lubridate`, `data360r`, `stringr`, `dplyr`, `zoo` (if they have not been previously installed, these packages are installed automatically when the function is called). OS agnostic (Linux, OS X, MS Windows).
Link to developer documentation and user manual	https://github.com/mtkurbucz/hdData360r/blob/main/README.md
Support email for questions	kurbucz.marcell@wigner.hu

1. Motivation and significance

The World Bank Group's GovData360 [1] and TCdata360 [2] platforms are widely employed in socio-economic research. The

* Correspondence to: Department of Computational Sciences, Wigner Research Centre for Physics, 29-33 Konkoly-Thege Miklós Street, Budapest, H-1121, Hungary.
E-mail address: kurbucz.marcell@wigner.hu.
https://doi.org/10.1016/j.softx.2022.101297

thousands of governance, trade, and competitiveness indicators they contain served as the basis for much research in the field of economic [3,4], developmental [5,6], and cultural [7,8] studies, coronavirus disease 2019-related research [9–11], and tourism [12,13], to state a few. In addition to the wide range of available indicators, the major advantage of the aforementioned platforms is that they contain data going back decades for all countries worldwide.

Although an R package called `data360r` [14] is available to facilitate access to the GovData360 and TCdata360 platforms, this package does not allow the simultaneous download of all available indicators. Furthermore, it does not allow the automatic replacement of missing values (e.g., with data from previous years), which would significantly facilitate the use of the compiled dataset. To fill these gaps and to support the rapidly growing [15] data-driven approach within social research, this study presents a novel R package named `hdData360r`. This package not only collects the up-to-date annual indicators from the aforementioned platforms but also preprocesses them for further analysis. This package is based on an improved version of a widely used R script [10,11,16] published together with the common database of COVID-19 reports and `data360` indicators (see Kurbucz, 2020 [17]).

The rest of this paper is organized as follows. Section 2 describes the `hdData360r` R package, as well as every related file. Section 3 presents the installation steps of the package. Section 4 gives an illustrative example for the application of the package. Finally, Section 5 provides an impact overview and concludes the paper.

2. Software description

The presented R package called `hdData360r` collects and preprocesses thousands of up-to-date annual governance, trade, and competitiveness indicators from World Bank Group platforms for all countries worldwide. It contains a function named `get_hdData360` that has one mandatory and two optional user-defined arguments. The mandatory argument (`firstYear`) specifies the first year of data collection, while the first optional argument (`impute`) can be used for missing value imputation. The `get_hdData360` function generates a list object that contains the panel data, the metadata of its indicators, the country data, and the information of the data generation process. The second optional argument (`writeTSV`) allows exporting this object into tab-separated value (`TSV`) files.

To obtain the GovData360 and TCdata360 indicators, as well as the country data and the metadata of the indicators, `data360r` (version: 1.0.8) R package [14] is applied. Other R packages used during the data generation process are: `lubridate` (version: 1.8.0) [18], `stringr` (version: 1.4.0) [19], `dplyr` (version: 1.0.9) [20], and `zoo` (version: 1.8-10) [21]. These packages are installed automatically when the function is called (if they have not been previously installed). The `hdData360r` R package with a sample dataset it generates is available publicly on GitHub [22] and Mendeley Data [23].

Properties of main files:

- **get_hdData360** (*get_hdData360.R*): The `hdData360r` R package contains a single function called `get_hdData360`. It has one mandatory and two optional user-defined arguments. The mandatory input parameter ($\texttt{firstYear} \in \{1970, 1971, \ldots, \text{“actual year”}\}$) defines the first year of the data collection. The second input parameter ($\texttt{impute} \in \{0, 1, \ldots, 30\}$) is optional, and it specifies that maximum how many years older data is used to replace missing values. That is, if the values of an indicator are (1, **NA**, **NA**) in three consecutive years, respectively, then, in the case of $\texttt{impute} = 1$, the missing value (**NA**) of the second year is imputed by the value of the first year (1, 1, **NA**). In the case of $\texttt{impute} = 2$, both missing values are imputed by the first year's value (1, 1, 1). By default, $\texttt{impute} = 0$, which results in no missing value imputation. The third, also optional, parameter ($\texttt{writeTSV} \in \{\texttt{FALSE}, \texttt{TRUE}\}$) allows exporting the generated object into tab-separated value (`TSV`) files. By default, its value is `FALSE`. The output list object has the following structure and contains the following variables:

  - `data`: Generated panel data.
    - `iso3` [*character*]: ISO 3166-1 alpha-3 (three-letter) country codes.
    - “`indicators`” [*character, integer, logical, numeric*]: Collected indicators. The indicator names are determined based on their identifiers and years.
  - `meta`: Metadata of indicators.
    - `id` [*integer*]: Identifier of indicators.
    - `name` [*character*]: Name of indicators.
    - `dataset` [*character*]: Name of the source dataset.
    - `valueType` [*character*]: Value type.
    - `datasetId` [*integer*]: Identifier of the source dataset.
    - `datasetLink` [*character*]: Link to the source dataset.
    - `defaultViz` [*character*]: Default visualization type.
    - `doNotUseViz` [*list*]: Visualization types that cannot be used.
    - `definition` [*character*]: Definition of indicators.
    - `units` [*character*]: Unit of indicators.
    - `subindicatorType` [*character*]: Type of the subindicator.
    - `timeframes` [*list*]: Time frame of the indicators.
    - `periodicity` [*character*]: Periodicity of the indicator.
    - `dateRange` [*character*]: Date range of indicators.
    - `site` [*character*]: Source repository (GovData360 or TCdata360).
  - `ctry`: Country data.
    - `id` [*character*]: Identifiers of countries.
    - `iso2` [*character*]: ISO 3166-1 alpha-2 (two-letter) country codes.
    - `iso3` [*character*]: ISO 3166-1 alpha-3 (three-letter) country codes.
    - `name` [*character*]: Name of countries.
    - `region` [*character*]: Region of countries.
    - `adminRegion`: Administrative region of countries.
    - `incomeLevel` [*character*]: Income level of countries.
    - `lendingType` [*character*]: Lending type.
    - `capitalCity` [*character*]: Capital of countries.
    - `geo` [*data.frame*]: Latitude and longitude of countries.
  - `info`: Information of the data generation process.
    - `timestamp` [*POSIXct*, *POSIXt*]: Timestamp of the data generation.
    - `firstYear` [*numeric*]: Mandatory parameter of the data generation.

* `impute` [*numeric*]: First optional parameter of the data generation.

- **Dataset generated by the `get_hdData360` function** (*hdData360r__‘‘firstYear’’_‘‘impute’’_“date”*): Directory with four tab-separated value (TSV) files that follow the structure of the output list object detailed above.

3. Installation

The `hdData360r` can be installed by using `devtools` R package as follows.

```
# install.packages("devtools")
# library(devtools)
install_github("mtkurbucz/hdData360r")
```

4. Illustrative examples

This section presents a simple example of using the generated dataset. In this example, we focus on the ratio of export and import value of various COVID-19 medical products in 2020. Although the World Health Organization (WHO) classified COVID-19 as a pandemic only on 11 March 2020, the export of related medical products was already restricted in January 2020 by the governments of more than 50 countries worldwide [24–26]. The indicators collected by the `hdData360r` package can help to gain a deeper understanding of changes in the global trade network of such products.

To demonstrate this, first, the BACI dataset [27], containing the bilateral trade flows for 200 countries, was collected from the CEPII website [28]. After that, the website of World Integrated Trade Solution [29] was used to identify the COVID-19 medical products (marked by six-digit codes, HS-6) and their categories. These are as follows:

A: Medical test kits (*HS-6: 300215, 382100, 382200, 902780*);
B: Disinfectants and sterilization products (*HS-6: 220710, 220890, 284700, 300490, 380894, 841920*);
C: Other medical consumables (*HS-6: 280440, 300510, 300590, 300670, 340111, 340120, 392329, 392690, 481890, 901831, 901832*);
D: Other medical devices and equipment (*HS-6: 732490, 841319, 901811, 901812, 901890, 902212, 902519, 902780, 902820*);
E: Other medical related goods (*HS-6: 731100, 761300, 842139, 940290*);
F: Oxygen therapy equipment and pulse oximeters (*HS-6: 901819, 901839, 901920, 902680*);
G: Protective garments (*HS-6: 392620, 401511, 401519, 401590, 481850, 611610, 621010, 621050, 621600, 630790, 650500, 900490, 902000*);
H: Vehicles (*HS-6: 870590, 871310, 871390*).

To visualize the relationship between the export–import ratio of each category and various governance, trade, and competitiveness indicators collected by the `get_hdData360` function, the `tabplot` (version: 1.4.1) [30] R package is used. This package provides table plots to explore and analyze large multivariate datasets. In our case, each column of this plot represents a medical product category, and each row represents a bin containing 100 indicators from GovData360 and TCdata360 platforms. Bars show the mean and the standard deviation of Spearman's correlations between the given product category and indicators contained in the bins. The last bar of the plot displays the ratio of the GovData360 and TCdata360 indicators for each bin. The table plot is illustrated in Fig. 1.

In absolute value, the highest Spearman's rank correlation coefficient ($\rho = 0.759$) was identified between the Global Innovation Index [31] and the export–import ratio of medical test kits. To illustrate how innovation capacity affects the international trade of different COVID-19 medical products, Fig. 2 shows the relationship of all product categories with this indicator.

As Fig. 2 shows that, in contrast to medical test kits, the export–import ratio of protective garments (e.g., gloves and masks) has only a slightly positive correlation with the Global Innovation Index ($\rho = 0.256$). The positive correlation observed for the other products is moderately strong. These results reflect that the more innovative countries are typically net exporters of healthcare products that have higher technological demands.

5. Impact and conclusion

The presented `hdData360r` R package mainly supports the rapidly growing data-driven approach within social research. It collects thousands of up-to-date annual governance, trade, and competitiveness indicators from the World Bank Group's GovData360 and TCdata360 platforms for all countries worldwide. The main advantage over the existing packages, such as the `data360r` Application Programming Interface (API) of the aforementioned platforms, is that it allows the simultaneous download of all available indicators. As for additional features, it is able to impute missing values with data from previous years, and optionally, it exports the generated dataset into tab-separated value (TSV) files.

The `hdData360r` package is based on an improved version of a widely used R script [10,11,16] published together with the common database of COVID-19 reports and `data360` indicators (see Kurbucz, 2020 [17]).

In conclusion, the value of the `hdData360r` function can be summarized as follows:

- The `hdData360r` R function collects thousands of up-to-date annual governance, trade, and competitiveness indicators.
- Optionally, it preprocesses the collected panel data.
- It also collects information about the countries, indicators, and the data generation process.
- Optionally, it exports the panel data with the additional information into tab-separated value (TSV) files.
- The generated dataset supports mainly the rapidly growing data-driven approach within social research.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marcell T. Kurbucz reports financial support was provided by Ministry for Culture and Innovation.

Data availability

I have shared a link to my data in the manuscript (https://data.mendeley.com/datasets/jwkk44trj3).

Acknowledgments

Supported by the ÚNKP-22-4-II-CORVINUS-55 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund, Hungary.

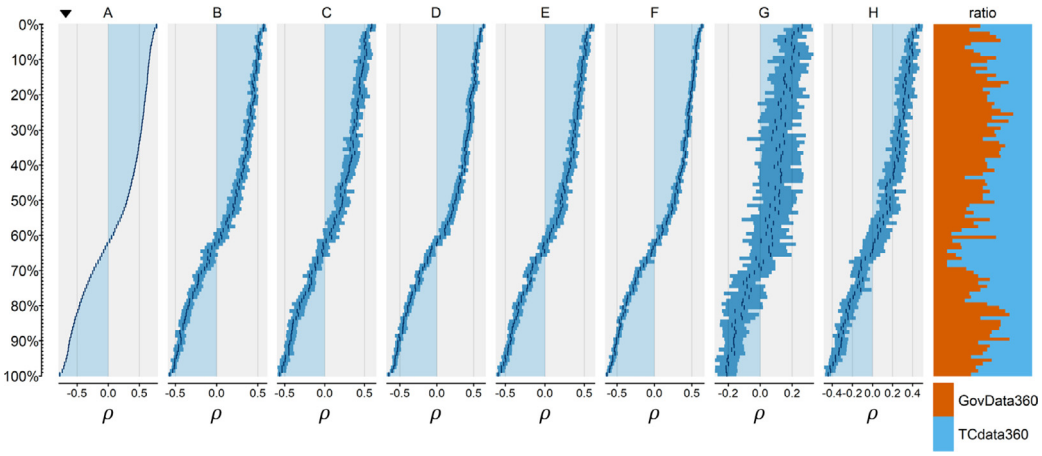

Fig. 1. Correlation between the export–import ratio of COVID-19 medical products and the innovation capacity in 2020
(*Remarks: Spearman's correlation coefficient is noted with ρ. The data is sorted by category A. Each bin (row) contains 100 indicators. The last bar displays the source ratio of bins.*).

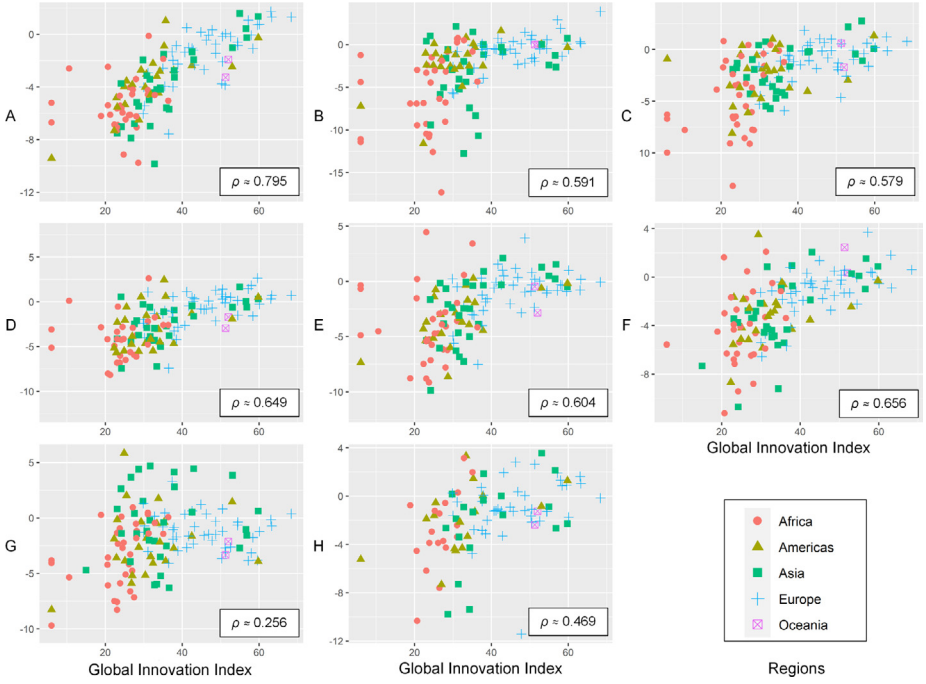

Fig. 2. Relationship between the export–import ratio of COVID-19 medical products and the innovation capacity in 2020
(*Remarks: Spearman's correlation coefficient is noted with ρ. Trade ratio is measured on a logarithmic scale.*).

References

[1] World Bank Group. GovData360. 2022, URL https://govdata360.worldbank.org/, (Accessed 03 December 2022).

[2] World Bank Group. TCdata360. 2022, URL https://tcdata360.worldbank.org/, (Accessed 03 December 2022).

[3] Echevarría CA, García-Enríquez J. The economic cost of the Arab Spring: the case of the Egyptian revolution. Empirical Econ 2020;59(3):1453–77.

[4] Kohler K, Stockhammer E. Growing differently? Financial cycles, austerity, and competitiveness in growth models since the Global Financial Crisis. Rev Int Polit Econ 2021;1–28.

[5] Alam M, Dappe MH, Melecky M, Goldblatt R. Wider economic benefits of transport corridors: Evidence from international development organizations. J Dev Econ 2022;102900.

[6] Munir M, Zakaria ZA, Baig AA, Mohamad MB. Development of global education index and establish relationship with human obesity through human development levels clustering. Int J Special Educ 2022;37(3).

[7] Ng SI, Lim XJ. Are Hofstede's and Schwartz's values frameworks equally predictive across contexts? Rev Brasileira GestÃO NegÓCios 2019;21:33–47.

[8] Bellido H, Marcén M, Morales M. The reverse gender gap in volunteer activities: Does culture matter? Sustainability 2021;13(12):6957.

[9] Salvador CE, Berg MK, Yu Q, San Martin A, Kitayama S. Relational mobility predicts faster spread of COVID-19: A 39-country study. Psychol Sci 2020;31(10):1236–44.

[10] Kurbucz MT, Katona AI, Lantos Z, Kosztyán ZT. The role of societal aspects in the formation of official COVID-19 reports: A data-driven analysis. Int J Environ Res Publ Health 2021;18(4):1505.

[11] Kurbucz MT. Modeling the social determinants of official COVID-19 reports in the early stages of the pandemic. J Appl Soc Sci 2022;16(1):356–63.

[12] Khan MYH, Islam ST, Hassan A. Factors influencing capital investment in the Bangladesh tourism industry. In: Tourism in Bangladesh: Investment and development perspectives. Springer; 2021, p. 63–78.

[13] Yang Y, Fan Y, Jiang L, Liu X. Search query and tourism forecasting during the pandemic: When and where can digital footprints be helpful as predictors? Anna Tourism Res 2022;93:103365.

[14] Ramin R, Onglao-Drilon P. data360r': Wrapper for 'TCdata360' and 'Govdata360' API. R Package Version 2020;1(8).

[15] Zhang J, Wang W, Xia F, Lin Y-R, Tong H. Data-driven computational social science: A survey. Big Data Res 2020;21:100145.

[16] Kosztyán ZT, Kurbucz MT, Katona AI. Network-based dimensionality reduction of high-dimensional, low-sample-size datasets. Knowl-Based Syst 2022;109180.
[17] Kurbucz MT. A joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of World Bank group platforms. Data Brief 2020;31:105881.
[18] Grolemund G, Wickham H. Dates and times made easy with lubridate. J Stat Softw 2011;40(3):1–25, URL https://Www.Jstatsoft.Org/V40/I03/.
[19] Wickham H, Wickham MH. Package 'stringr'. 2019, URL https://cran.r-project.org/web/packages/stringr/index.html.
[20] Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 0.4. 3. R Found. Stat. Comput., Vienna 2015. URL https://CRAN.R-project.org/package=dplyr.
[21] Zeileis A, Grothendieck G. Zoo: S3 infrastructure for regular and irregular time series. J Stat Softw 2005;14(6):1–27. http://dx.doi.org/10.18637/jss.v014.i06.
[22] Kurbucz MT. hdData360r: A high-dimensional panel data compiler for governance, trade, and competitiveness indicators of World Bank Group platforms. 2022, GitHub, URL https://Github.Com/Mtkurbucz/HdData360r/.
[23] Kurbucz MT. Panel data generated from governance, trade, and competitiveness indicators of World Bank Group platforms. 2022, Mendeley Data, URL https://Data.Mendeley.Com/Datasets/Jwkk44trj3/.
[24] Alert GT. Tackling COVID-19 together: The trade policy dimension. Technical report, University of St. Gallen, Switzerland; 2020, URL https://www.globaltradealert.org/reports/51.
[25] Bown CP. COVID-19: Demand spikes, export restrictions, and quality concerns imperil poor country access to medical supplies. COVID-19 and Trade Policy: Why Turning Inward Won'T Work 2020;31–48, VoxEU.org eBook, CEPR Press London.
[26] Grassia M, Mangioni G, Schiavo S, Traverso S. (Unintended) Consequences of export restrictions on medical goods during the COVID-19 pandemic. J Complex Netw 2022;10(1).
[27] Gaulier G, Zignago S. Baci: International trade database at the product-level. 2010.
[28] CEPII. BACI. 2022, URL http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37. (Accessed 3 December 2022).
[29] World Integrated Trade Solution. COVID-19 medical products. 2022, URL https://wits.worldbank.org/trade/covid-19-medical-products.aspx, (Accessed 03 December 2022).
[30] Tennekes M, de Jonge E. Package 'tabplot'. 2019, R Package, URL https://Cran.Microsoft.Com/Snapshot/2018-04-09/Web/Packages/Tabplot/Tabplot.Pdf.
[31] World Intellectual Property Organization. Global Innovation Index. 2022, URL https://www.globalinnovationindex.org/, (Accessed 03 December 2022).