MethodsX 8 (2021) 101567

Contents lists available at ScienceDirect

MethodsX

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

Method Article

Automated cleansing and harmonization of international trade data

Sandra Oliveira*, César Capinha, Jorge Rocha

Centre for Geographical Studies and Associated Laboratory TERRA, Institute of Geography and Spatial Planning, Universidade de Lisboa, Lisbon, Portugal

A B S T R A C T

Large volumes of data are becoming increasingly available and can be very valuable for the analysis of different phenomena. These data can originate from multiple sources and be recorded in diverse formats, requiring preliminary scrutiny in order to be further used in scientific analyses. This first crucial phase of filtering and cleansing data is usually a cumbersome and time-consuming task, but automated routines can be developed to help researchers. A routine created with the R language is here presented, to screen, harmonize and aggregate international trade data, representing the trade flows between countries for specific products, in a timeframe that covers monthly flows for at least 15 years for most countries. The R script implementing these routines is provided, being easily adapted to other datasets with similar issues.

- A step-by-step procedure for cleansing and harmonizing international trade data, using R programming language, is presented
- Automated routines are very effective in obtaining robust and filtered data inputs to integrate in scientific models
- Spatial and temporal patterns of worldwide trade relations can be explored to enhance our understanding of various associated phenomena

A R T I C L E I N F O

Method name: Cleansing and harmonization of international trade data
Keywords: Automated screening, Data harmonization, Time-series analysis, R software
Article history: Received 9 September 2021; Accepted 30 October 2021; Available online 2 November 2021

* Corresponding author.
E-mail address: sandra.oliveira1@campus.ul.pt (S. Oliveira).

https://doi.org/10.1016/j.mex.2021.101567

Specifications Table

Subject Area:	Environmental Science
More specific subject area:	Spatial and time-series data analysis
Method name:	Cleansing and harmonization of international trade data
Name and reference of original method:	*Not applicable*
Resource availability:	R script available as supplementary material

Data acquisition

Access to statistical data concerning the international trade of goods (imports) was obtained from the International Trade Centre (ITC) [1]. The ITC enlarged the digital open access of its database for several months in 2020, providing monthly statistics of imported goods by country free of charge, via their website [1]. The available information for specific commodities potentially associated with the accidental spread of Aedes spp. mosquitoes, specifically tyres (new and used) and live plants, was collected for 79 countries.

The access to this database faced several constraints, in particular the limited amount of data that could be downloaded in each web interaction and their saving format. Due to these limitations, the data retrieved for each country was composed of a set of separate files, the total number depending on the years available, and the last column of a file was automatically assigned as the first column of the following file, creating duplicates. The structure of the files for one single country could also differ, specifically related to two issues: i) the number of rows in each file was different, because the number of exporter countries varied over time; ii) the number and position of month/year columns could differ, because other information was sometimes added besides the date (e.g. units). In addition, the records for each individual country could be presented in different units (for example kilograms, tons or number of individual units), in some cases with mixed quantities for the same date, which could result in a missing (zero) or partial value for the total amount of imports, calculated as a "World" row that was provided in the original data alongside each specific exporter country.

Method details

A cleansing and harmonization procedure to transform and organize the trade data obtained from ITC was developed, using R scripting tools and data wrangling packages [2–8]. The proposed methodology (Fig. 1) was applied to each country and product individually, storing the results in specific folders created along the screening process, to maintain the database organized for further use.

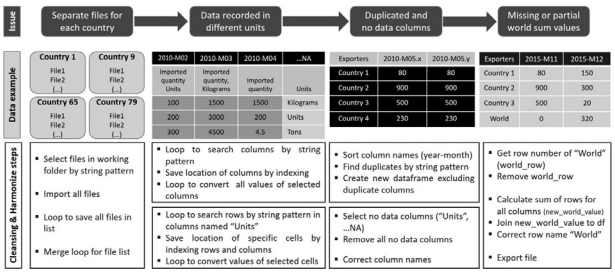

Fig. 1. Scheme of the overall methodology developed with R programming language for the cleansing and harmonization of the international trade data, including the major issues, data examples and processing steps.

Separate files for each country

The first issue to overcome was the data being saved in separate files for each product and country, due to table size limitations in the downloading process (Fig. 1). In each web interaction, only a restricted number of date columns could be downloaded (n=20), which recorded the year and month of the trade flow. For most countries data was available between 2004 and 2019, but in some cases the available period was shorter, therefore the number of downloaded files could also differ between countries. All the corresponding files (for product and country) were imported to R software and stored as a list, based on string pattern search within the working folder, and were afterwards merged in a loop by the name of the countries from where the product is imported (Exporters). The result is a single file with the amount of imports per year-month in columns, and the name of the multiple Exporter countries in rows.

Data recorded in different units

The second issue arose from the registration of the imports in different units, depending on the Exporter country, the date or both. As such, some of the columns did not correspond to import amounts (hereby called values), but instead to the description of the unit assigned to the value that was recorded in the previous column (Fig. 1), such as kilograms or individual units of the product. Moreover, different units could be recorded in the same column. These characteristics of the original data required a harmonization step that consisted in searching, within the merged dataframe, the location of the values recorded in different units and subsequently converting them to the same weight unit (in this case, to tons). The search was followed by an indexing function, based on the position of the corresponding columns and rows, since this helps define pointers to where the values are stored within a dataframe. This had to be implemented in two different ways:

1) When the entire column recorded the same weight unit but different from tons, the search was based on a full string pattern that was found in the last row of the merged dataframe, such as “Imported quantity, Units”. The index position of these columns was then used to convert simultaneously all the values of the columns that matched the pattern search criteria.
2) In the case of the columns where the values were recorded in multiple weight units, the search for each specific value that required conversion was based on the indexing of the corresponding column and row where each value was located.

To automate this step, the search, indexing and conversion of the values were done with iterating functions. To ensure the applicability of the script for different countries and the execution of the whole procedure without obstacles, and due to the possible different formats and structure of the original data per country and product, each of the looping functions was only run if the appropriate conditions were met, otherwise the execution of the script would proceed to the following step.

Duplicated and no data columns

The third main issue was related to the duplication of columns, resulting from the merge of the original files, and to the existence of columns that presented the description of the type of unit assigned, instead of import amounts. This was handled through the cleaning and sorting of columns, using iterations for string pattern search and replacement based on column names, the selection of the columns which matched specific criteria and the subsequent filtering of the merged dataframe by excluding the selected columns. The remaining columns were then sorted by ascending order of year-month, creating a structured time-series of product imports for each specific country.

Missing or partial world sum values

The final issue was associated with the mixed types of units that were recorded in the same column, which resulted in missing or partial values being provided for the row that collected the “World” value (sum of all imports) for the product and country. As such, the original row named

"World" in the column of Exporters was replaced by the updated sum of all rows for each year-month column, and this was placed as the last row of the dataframe, overtaking the strict alphabetical order of Exporters names that the previous steps had automatically implemented.

Conclusions

The data screening procedure here presented was able to overcome the several issues found in the data acquired. The functions applied are easily accessible from R software and corresponding packages, and it can be implemented by different users, even those with limited experience in programming tools. For each product and country, the whole script runs in about 1 minute with computation capabilities up to 16 Gb RAM, and it is possible to adjust it to run for several countries simultaneously. The screened data contains a fully harmonized time-series database with the amount (weight in tons) of imports regarding live plants and tyres, most being available between 2004 and 2019. The monthly data that became accessible after the implementation of this procedure, allows for the analysis of seasonal and annual trends of trade flows between countries and regions, which can be explored in association with other phenomena, such as the potential spread of species that disseminate vector-borne diseases [9–10]. The base script that assembles the whole procedure is available as a supplementary file and can be adapted to other data with similar issues.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financed by national funds through FCT—Portuguese Foundation for Science and Technology, I.P., under the framework of the Project "TRIAD—health Risk and social vulnerability to Arboviral Diseases in mainland Portugal" [PTDC/GES -OUT/30210/2017] and by the Research Unit UIDB/00295/2020 and UIDP/00295/2020. CC was funded through FCT, I.P., under the programme of 'Stimulus of Scientific Employment—Individual Support' within the contract 'CEECIND/02037/2017'.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2021.101567.

References

[1] ITC, International Trade Centre. Trade statistics for international business development. Available from www.trademap.org [Retrieved in June-July 2020]
[2] R Core TeamR: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2020 URL https://www.R-project.org/ .
[3] Wickham, H.; François, R.; Henry, L.; Müller, K. (2020). dplyr: a Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
[4] Wickham, H. (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr
[5] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D.A. McGowan, R. François, … H. Yutani, Welcome to the Tidyverse, J. Open Source Softw. 4 (43) (2019) 1686, doi:10.21105/joss.01686.
[6] Wickham, H.; Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
[7] Dragulescu, A.; Arendt, C. (2020). xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.5. https://CRAN.R-project.org/package=xlsx
[8] Microsoft and Steve Weston (2020). foreach: Provides Foreach Looping Construct. R package version 1.5.1. https://CRAN.R-project.org/package=foreach
[9] P. Reiter, Aedes albopictus and the world trade in used tires, 1988-1995: the shape of things to come? J. Am. Mosq. Control Assoc. 14 (1) (1998) 83–94.
[10] P.E. Hulme, Unwelcome exchange: International trade as a direct and indirect driver of biological invasions worldwide, One Earth 4 (5) (2021) 666–679.