Article

Daniel Antal, Marta Kolczynska

7/6/2020

The increasing availability of microdata available as open data creates new opportunities for ex-post harmonization and joint exploitation of these resources. We introduce the surveyharmonization and eurobarometer packages that help the reproducible ex-post harmonization of survey data stored in various, individual files with rich metadata. Our examples are related to joining data from SPSS files of various Eurobarometer survey responses spanning over 25 years, but our tools are useful to make this work.

The Eurobarometer case is particularly interesting, because each wave of the Eurobarometer survey is n fact a combination of 27-30 nationally representative survey conducted in about 20 languages based on a generic English and French questionnaire. They contain trend variables, which are asked repeatedly in many waves, so only joining the same variables of two Eurobarometer waves, for example, on attitudes to trust in institutions, requires a harmonization of about 60 complex data assets. Much of this harmonization work is done ex ante, and it is mainly outside of the scope of our interest. While conducting the fieldwork of these surveys and processing the interview data is fully harmonized within each wave, there are various ex ante harmonization solutions are applied in various waves, especially ones that took place far away in time, for example, in 1995 and 2019. When we are ex post joining data from different survey waves, we have to understand and partly reconstruct ex ante harmonization steps.

Ex-post (or retrospective) data harmonization refers to procedures applied to already collected data to improve the comparability and inferential equivalence of measures collected by different studies. In practice, ex-post harmonization means a lot of recoding, i.e. to make sure that codes are understood exactly the same way across all data resources.

Reproducible recoding can be in form of code in some programming language, where running the code performs all the data transformations. When the number of recodes is large, the code quickly becomes long and complicated and difficult to follow for anyone but the code’s creator. We tried to create a workflow that is fully reproducible and automatically documented.

In simple words, when we are creating a data panel from two (or more) Eurobarometer surveys stored in separate SPSS files, we are aiming to solve the following problems:

– [Importing data](#import) from proprietary statistical software and creating a faithful, practical representation of their data and metadata contents for later harmonization and join.

– [Ex-post harmonization](#expost)

* Define a common [taxonomy](#taxonomy), probably based on a thesaurus for consistentcy.
* [harmonizing variable names](#harmonize-var-names), or “variable labels” in an SPSS context, making sure that variables that represent exactly the same information have exactly the same, programmatically usable variable names; dissimilar variables have dissimilar names.
* [Harmonizing value labels](#harmonize-labels) of categorical variables, making sure that the same labels, such as ‘good’, ‘mediocre’, ‘bad’ always have the same representation in our new data object.
* Harmonizing the treatment of [missing values](#harmonize-missing), carefully documenting the differences of ‘missingness’ (inapplicable items, items that were not collected, etc.), because during analysis they may require different interpretation.

– Creating a [joint data table](#datajoin) where each variable that represents the same class and metadata attributes; documenting all the changes and Providing methods that re-export the data files to SPSS, Stata, SAS.

* Creating a consistent, [unique identifier](#uniqid) for each observation.
* Create a full [documentation](#documentation) of conversions, recoding for reproducability.
* Provide methods to export the harmonized, joined data to proprietary formats for [analysis outside R](#exportmethods)
* If the user wishes to prepare reproducible analysis in R, provide [methods](#rmethods) to convert the data to base types of numeric and factor, the usual inputs of statistical functions.

Our workflow generally fits into the tidyverse.

## Import files

Complex microdata, such as primary survey data, is often stored in some statistical file format. In our article, we use examples for SPSS’s .sav format, but similar problems are present with other proprietary file formats. They usually do not only represent the data, but rich metadata that provides guidance for the analyst who wants to use them – for example, which observations should be treated as missing, or what is the valid range of a variable.

In our case, we are dealing with highly structured survey data that contain a lot of metadata, partly about the ex ante harmonization process that is vital for using the data. For example, an archived Eurobarometer wave file contains recently more than 28 subsamples (for each EU member state, two for Germany, two for the former member state United Kingdom, and often for EEA or prospective member state countries.) The way these subsamples were merged together and handled must be preserved.

We fully rely in importing on two packages, [haven](https://haven.tidyverse.org/) and [labelled](https://cran.r-project.org/web/packages/labelled/vignettes/intro_labelled.html) from the tidyverse. They have created classes and methods that give a faithful representation of imported statistical software files, but they need to be amended for reproducibility and harmonization information for later joins. These steps are likely to be present when a researcher is joining any two statistical software files, not only survey microdata files.

## Ex-post harmonization

maybe you code write a bit about ex-post harmonizaton here. This can also lead out to your other work or to a later article about describing a trend file that we create.

### Thesaurus, taxonomy

### Harmonization of variable names

We should keep an eye on [snakecase](#https://github.com/Tazinho/snakecase) and maybe janitor, too.

### Harmonization of categories

We should keep an eye on [forcats](#https://forcats.tidyverse.org/)

### Harmonization of missing values

## Joining the data

At this stage, we are ready to have rich data tables with many metadata attributes that are ready to be joined in R. We have kept all information that is necessary for the later step, i.e. the actual statistical analysis of the data in a way that it still can be done in SPSS, Stata, or in R.

* attributes are harmonized and joined
* data is harmonized and join

At this point the new data table is ready for analysis. We provide some methods for exporting them (back) to SPSS or Stata, or flat tables, before turning to analyzing the data in R.

### Unique identification

### Documenting the process

### Methods for exporting

* exporting to SPSS
* exporting to Stata
* exporting to flat files

### Methods for analyzing in R