EES2019 Stacking Process

Auxiliary Data Frames Enhancement

Giuseppe Carteny

23.08.2021

Contents

1 Introduction: The following tasks

As mentioned during our last call (23.08.2021), the following tasks for creating the EES 2019 voter study stacked data matrix (SDM) are the following:

1. Reviewing the scripts you sent me last week, merging the two auxiliary datasets that we used for selecting the relevant parties in each EU party system (namely the EES2019 Codebook and the 2019 European elections results)[1];
2. Creating the codebook of the dataset.
3. Estimating the generic variables of the SDM, as explained in our first call, and summarized in the presentation that I sent you at the beginning of our job.

This short document consists in a short tutorial for facing the step (1). Once completed it other I will provide the tutorials for the following two steps.

2 Why Enhancing our Auxiliary Data Frames

After reviewing the scripts for stacking the EES 2019 original data frame I noticed that the current workflow can be implemented. In particular I noticed that using two distinct dataframes, *that differ just a limited set of information*, for selecting the relevant parties consists a sub-optimal workflow. And this workflow can be enhanced by **merging the two dataframes**.
Although this process implies that we have to review the job that has been done few days ago, I think that this step is required for the following reasons:

1. It will make the relevant parties selection process more transparent and understandable, thus contributing to the **reproducibility** and **replicability** of our workflow. Since I do believe (as many others) that reproducibility and replicability are core concepts of scientific research in any field, I think that this already represents a good reason to enhance our workflow;
2. It will allow us to create **a first version of a dataframe that bridges the EES 2019 voter study and other data sources** (not only the European Parliament elections data, as in this case). This would be *really* useful (a) for the ProConEU project and perhaps (b) for other researchers that might be interested in using our data. I think that we might publish this 'bridge dataset' when the new version of the EES dataset will be published on Gesis, but this is just my own speculation that I need to discuss with other members of the ProConEU project.
3. It will allow to make the SDM creation **less dependent on manual coding**, more interactive, giving the opportunity to create SDM according to different criteria just switching a few objects in the R workflow. In perspective, this might allow us to create an app/interface for creating the SDM without reviewing the whole script.

[1] As you will notice, the EP 2019 election results dataset now contains not only the share of votes obtained by each party but also the *seats* obtained by some of the parties participating to said elections.

3 How to Enhance our Auxiliary Data Frames

Once justified the change in the current workflow let's talk about *how* to do it. I must say, as I already told you, that this task will be really (*really*) boring, because unfortunately we must resort to **manual coding**. I tried, as I always try, to avoid it since it is prone to error, mistakes, it is very painful to debug once you face a problem,... but unfortunately the two datasets (the EES 2019 voter study codebook and the 2019 European Parliament election results) do not share any common variable, and the existing variables do not follow similar coding patterns that might allow to resort to more computationally refined methods. I am sorry, but unfortunately the world of data sometimes can be very sad (and, again, boring).

The good news, however, is that coding 6/7 countries it is definitely not an enormous job, at least compared to manually coding 28 countries. So let's see how to do it first looking at the new workflow, and then looking at new scripts that need to be created and those that need to be reviewed.

3.1 General workflow

The workflow for enhancing the auxiliary data frames is similar to the one that we have already developed for stacking the original variables of the EES 2019 voter study. A new script named 'EES2019_cdbk_enh' will (1) source our two auxiliary data frames, and then (2) it will source a set of country-specific scripts merging the EES2019 codebook and the EP results for each EES 2019 voter study sample. Finally (3) the script will pool together these scripts, binding them in a single, enhanced codebook. After this passage the stacking process will follow almost the same workflow that we developed earlier, although some adjustments will be required

3.2 Country-specific codebook scripts

The country specific scripts for creating the enhanced codebook should follow a simple structure. In the following lines I will present such structure using the EES 2019 codebook and the 2019 EP elections results for Italy.

For working on the scripts first run the current version of the 'EES2019_stack.R' script **until line 42**. This will allow to load the R packages and the two auxiliary data frames, plus the EES 2019 original dataframe. After this passage we can start creating our country-specific dataset.

First, banally, we must select data referring to the country that we are interested in.

```
EES2019_cdbk_it <-
  EES2019_cdbk %>%
  filter(countryshort=='IT')

EP2019_it <-
  EP2019 %>%
  filter(countryshort=='IT')
```

Then we must print on our console the two datasets and check how to create a common variable on the EP2019 result that will allow us to merge the data of interest (in our case, party vote shares and seats) with the data already in the EES2019 codebook.

```
# Print the two country-specific auxiliary dataframes for coding purposes,
# but mute them once the coding process is completed.

EES2019_cdbk_it %>%
  dplyr::select(partyname, partyname_eng, Q7)
```

```
## # A tibble: 15 x 3
##    partyname                                 partyname_eng                    Q7
##    <chr>                                     <chr>                         <int>
##  1 "Partito Democratico (PD)"                Democratic Party               1501
##  2 "Forza Italia (Fi)"                       Go Italy                       1502
##  3 "Lega Salvini Premier"                    Northern League                1503
##  4 "Movimento 5 Stelle (MCS)"                Five Star Movement             1504
##  5 "Sinistra"                                Italian Left                   1505
##  6 "\"\"+Europa\"\""                         More Europe (+Europa)          1506
##  7 "Fratelli d’Italia - Centrodestra Nazion~ Brothers of Italy - National~  1507
##  8 "FEDERAZIONE DEI VERDI"                   Federation of the Greens       1508
##  9 "Sudtiroler Volkspartei (Partito popolar~ South Tyrol People’s Party     1509
## 10 "POPOLARI PER L’ITALIA"                   Populars for Italy             1510
## 11 "Partito Comunista"                       Communist Party                1511
## 12 "Forza Nuova"                             New orice                      1512
## 13 "Casa Pound"                              CasaPound Italy-United Right   1513
## 14 "Noi con L’Italia (Udc)_"                 Us with Italy_UDC                NA
## 15 "Italia Europa Insieme"                   Italy Europe Together            NA
```

```
EP2019_it %>%
  dplyr::select(partyname, partyname_eng, partyid)
```

```
## # A tibble: 10 x 3
##    partyname             partyname_eng                                    partyid
##    <chr>                 <chr>                                            <chr>
##  1 LN                    Lega Salvini Premier                             IT01
##  2 PD                    Partito Democratico (con Siamo Europei)          IT02
##  3 FI                    Forza Italia                                     IT03
##  4 FDI                   Fratelli d’Italia                                IT04
##  5 M5S                   Movimento Cinque Stelle                          IT05
##  6 Coal +E (+E + IC + P~ Coalition +Europa (+ Europa - Italia in Comune~ IT06
##  7 Coal La Sinistra (SI~ Coalition La Sinistra (Sinistra italiana + Rif~ IT07
##  8 SVP                   Südtiroler Volkspartei (Partito popolare sudti~ IT08
##  9 FdV (Verdi+Possibile~ Coalition Federazione dei Verdi (Verdi + Possi~ IT09
## 10 Other parties         Other parties                                   IT90
```

We can merge the two datasets in several ways. Nonetheless, since in our stacking process we rely on the identification codes of the original vote choice variable of the 2019 EES voter study (Q7; See the EES2019 Codebook) we can rely again on this variable. Then, the variable in the EP results that seems more suitable for our coding is `partyid`.

Consequently, for merging the two datasets we must create a new `Q7` variable in the EP2019 data frame, as it follows. We can do it using the `mutate()` and `case_when()` functions.
In order to avoid problems later on, it is important to remove the partyid value referring to the 'Other parties' category. Normally, this value consists in the abbreviation of the country ('IT' for the Italian sample) combined with 90.

```
EP2019_it %<>%
  filter(partyid!='IT90') %>%
  mutate(Q7 = case_when(partyid=='IT01' ~ as.integer(1503),
                        partyid=='IT02' ~ as.integer(1501),
                        partyid=='IT03' ~ as.integer(1502),
                        partyid=='IT04' ~ as.integer(1507),
                        partyid=='IT05' ~ as.integer(1504),
                        partyid=='IT06' ~ as.integer(1506),
                        partyid=='IT07' ~ as.integer(1505),
                        partyid=='IT08' ~ as.integer(1509),
                        partyid=='IT09' ~ as.integer(1508),
                        T~NA_integer_))
```

Once created the variable then we just have to merge the two datasets, selecting the variables of interest (namely EP2019 party vote shares and seats) and clean our environment.

```
EES2019_it_enhcdbk <-
  left_join(EES2019_cdbk_it,
            EP2019_it %>% dplyr::select(Q7, votesh, seats),
            by = 'Q7')

rm(list=ls(pattern='_it$'))
```

As you can see below now the EES2019 codebook has two new columns, referring to party vote shares and seats.

```
EES2019_it_enhcdbk %>%
  dplyr::select(partyname, partyname_eng, Q7, votesh, seats)
```

```
## # A tibble: 15 x 5
##    partyname                        partyname_eng               Q7  votesh seats
##    <chr>                            <chr>                    <int>   <dbl> <int>
##  1 "Partito Democratico (PD)"       Democratic Party          1501  0.227     19
##  2 "Forza Italia (Fi)"              Go Italy                  1502  0.0878     6
##  3 "Lega Salvini Premier"           Northern League           1503  0.343     28
##  4 "Movimento 5 Stelle (MCS)"       Five Star Movement        1504  0.171     14
##  5 "Sinistra"                       Italian Left              1505  0.0175     0
##  6 "\"\"+Europa\"\""                More Europe (+Europa)     1506  0.0311     0
##  7 "Fratelli d’Italia - Centrodest~ Brothers of Italy - Nat~  1507  0.0644     5
##  8 "FEDERAZIONE DEI VERDI"          Federation of the Greens  1508  0.0232     0
##  9 "Sudtiroler Volkspartei (Partit~ South Tyrol People’s Pa~  1509  0.0053     1
## 10 "POPOLARI PER L’ITALIA"          Populars for Italy        1510 NA         NA
## 11 "Partito Comunista"              Communist Party           1511 NA         NA
## 12 "Forza Nuova"                    New orice                 1512 NA         NA
## 13 "Casa Pound"                     CasaPound Italy-United ~  1513 NA         NA
## 14 "Noi con L’Italia (Udc)_"        Us with Italy_UDC           NA NA         NA
## 15 "Italia Europa Insieme"          Italy Europe Together       NA NA         NA
```

Once the script is completed then the next step will be launching again the main stacking script (namely, the ‘EES2019_it_stack.R’) until line 42. If everything works fine you should have in your environment an object called `EES2019_cdbk` that includes the data you just created in your country-specific script, plus the already existing ones.

3.3 Reviewing the country-specific stacking scripts

Once the new codebook is concluded then what we must review the scripts that we created earlier for stacking the original EES 2019 voter study variables. The reviewing process is rather straightforward.

First we change the ‘Filter the codebook and EP elections data’ section, changin also the title of said section in ‘Filter the codebook data’.

```
EES2019_cdbk_it <-
  EES2019_cdbk %>%
  filter(countryshort=='IT')
```

Then we must change the ‘Choose the relevant parties’ section. Note that the ‘votes_crit’ now includes also the `seats` variable, and that the code mutates the 0 values of the latter in `NA` values. This will allow, later on, to change our criteria more easily, but for now please just review the scripts without changing the criteria used earlier.

```
ptv_crit <-
  EES2019_cdbk_it %>%
  dplyr::select(partyname, Q10_PTV)
```

```
votes_crit <-
  EES2019_cdbk_it %>%
  mutate(seats = case_when(seats==as.integer(0) ~ NA_integer_, T~seats)) %>%
  dplyr::select(partyname, votesh, seats)

party <-
  EES2019_cdbk_it %>%
  dplyr::select(partyname, Q10_PTV, Q7) %>%
  na.omit() %>%
  .$Q7
```

If everything has been done properly then you might be able to run the country-specific script without any problem.

3.4 Problematic cases

In most cases both the creation of the new codebook scripts and the review of the former scripts should be relatively easy. Nonetheless there are some cases that are anything but straightforward, for several reasons.

First, we might face situations in which the party names are not properly encoded (and unfortunately cannot be encoded all at once) and this might create problems for properly identify the parties and thus creating the briding variable (the case below refers to the Cypriot sample):

```
## # A tibble: 15 x 3
##   partyname                          partyname_eng                         Q7
##   <chr>                              <chr>                              <int>
## 1 ???? (?????????? ????? ??????????? ?~ Progressive Party of the Working ~   501
## 2 ???? (???????????? ??????????)       Democratic Rally                     502
## 3 ???? (??????????? ?????)             Democratic Party                     503
## 4 ???? (?????? ??????????????? ????)   United Democratic Union of Centre    504
## 5 ?????? ?????????                     Ecological and Environmental Move~    NA
## # ... with 10 more rows
```

```
## # A tibble: 8 x 3
##   partyname  partyname_eng                                                partyid
##   <chr>      <chr>                                                        <chr>
## 1 <U+0394><U+0397>S<U+03A5>/DISY  <U+0394><U+03B7>µ<U+03BF><U+03BA><U+03C1>at<U+03B9><U+03BA><U+03CC
## 2 <U+0391><U+039A><U+0395><U+039B>/<U+0391><U+039A><U+0395>L  <U+0391><U+03BD><U+03BF><U+03C1><U+03B
## 3 <U+0394><U+0397><U+039A><U+039F>/DIKO  <U+0394><U+03B7>µ<U+03BF><U+03BA><U+03C1>at<U+03B9><U+03BA>
## 4 <U+0395><U+0394><U+0395><U+039A>/ EDEK <U+0395><U+0394><U+0395><U+039A> <U+039A><U+03AF><U+03BD><U
## 5 <U+0395><U+039B><U+0391><U+039C>/ELAM  <U+0395><U+03B8><U+03BD><U+03B9><U+03BA><U+03CC> <U+039B>a<U
## # ... with 3 more rows
```

Second, we might face situations with coalitions rather than parties, such as the Polish case showed below:

```
## # A tibble: 13 x 3
##    partyname                       partyname_eng                              Q7
##    <chr>                           <chr>                                   <int>
##  1 Platforma Obywatelska (PO)      Civic Platform                             NA
##  2 Polskie Stronnictwo Ludowe (PS~ Polish People’s Party                      NA
##  3 Sojusz Lewicy Demokratycznej (~ Democratic Left Alliance                   NA
##  4 Prawo i Sprawiedliwo?? (PIS)    Law and Justice                          2104
##  5 Kukiz’15                        Kukiz’15                                 2106
##  6 Wiosna Roberta Biedronia        Spring                                   2102
##  7 Razem                           Poland Together                          2105
##  8 Koalicja Europejska PO PSL SLD~ European Coalition                       2103
##  9 Konfederacja Korwin Braun Liro~ Coalition for the Renewal of the Repub~  2101
## 10 Polska Fair Play bezpartyjni G~ Poland Fair Play (PFP)                   2107
## 11 Ruch Prawdziwa Europa - Europa~ <NA>                                     2108
## 12 Polexit-Koalicja(P-K)           Coalition for the Renewal of the Repub~  2109
## 13 Nowoczesna Ryszarda Petru       Modern                                     NA
```

```
## # A tibble: 8 x 3
##   partyname              partyname_eng                                   partyid
##   <chr>                  <chr>                                           <chr>
## 1 PiS                    "Prawo i Sprawiedliwosc"                        PL01
## 2 Konfederacja           "Konfederacja KORWiN Braun Liroy Narodowcy"     PL02
## 3 Kukiz’15               "Kukiz’15"                                      PL03
## 4 Wiosna                 "Wiosna Roberta Biedronia"                      PL04
## 5 Coal KE (PO + PSL + S~ "Coalition Koalicja Europejska (Platforma Obyw~ PL05
## 6 Polska Fair Play       "Polska Fair Play Bezpartyjni Gwiazdowski "     PL08
## 7 Coal Lewica Razem (PR~ "Coalition Lewica Razem (Partia Razem + Unia P~ PL09
## 8 Other parties          "Other parties"                                 PL90
```

Unfortunately, there are not straightforward/automatic/general solutions for such issues. Thus, I suggest to contact me that when you face such situations, sending me the scripts, in order to make decisions case by case.

4 Who Does What

For accomplish the tasks discussed above I suggested to work on the same countries and scripts that we have been working with in the previous weeks.

- **Willie**: Denmark, Estonia, Germany, Luxembourg, Malta, Spain, United Kingdom;
- **Julian**: Czech Rep., Finland, Hungary, Lithuania, Slovakia, Poland, Sweden;
- **Matthias**: Austria, France, Ireland, Latvia, Portugal, Romania, Slovenia.

As before, I will take care of the remaining ones (namely: Belgium, Bulgaria, Croatia, Cyprus, Greece, Netherlands). However, as highlighted above, please contact me whenever you face any difficulty especially those summarised in the previous section (Sect. 3.4).

5 Deadlines

I think that this (boring) job should be completed before the end of the week (**27.08.2021**), in order to start with the following steps starting from next Monday (**30.08.2021**). In the case in which you will finish your job before the deadline then we will start our following tasks earlier (see Sect. 1).