EU-SILC Tools: eusilcpanel_2020 - first computational steps towards a cumulative sample based on the EU-SILC longitudinal datasets; Update

Borst, Marwin; Wirth, Heike

gesis Leibniz-Institut für Sozialwissenschaften

GESIS Papers

2022|10

EU-SILC Tools: eusilcpanel_2020

First computational steps towards a cumulative sample based on the EU-SILC longitudinal datasets

– Update –

Marwin Borst & Heike Wirth

GESIS Papers 2022|**10**

EU-SILC Tools: eusilcpanel_2020

First computational steps towards a cumulative sample based on the EU-SILC longitudinal datasets

– Update –

Marwin Borst & Heike Wirth

GESIS – Leibniz-Institut für Sozialwissenschaften 2022

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 269
Telefax: (0621) 1246 - 100
E-Mail: gml@gesis.org

Marwin Borst, ex-Department of Statistics, University of Rome "La Sapienza"
E-Mail: marwin.borst@live.it

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Abstract

The European Union Statistics on Income and Living Conditions (EU-SILC) covers a wide array of variables collected from households by the Member States. Among others, EU-SILC contains panel data that follows a rotational design. Each year, Eurostat publishes a series of separate datasets covering only up to 4 years, even though it has been collecting data since 2003. “eusilcpanel” is a script written by Marwin Borst (download: https://www.gesis.org/gml/european-microdata/eu-silc/) in the form of a Stata package (eusilcpanel.ado; eusilcpanel.sthlp; totalpopulation.dta), that is able to merge these chunks of data into one cumulative dataset (separately for the D-,H-,R-, and P-data). The script makes the EU-SILC panel more accessible to researchers in the vast majority of cases, but it can’t deal with data from all countries.

eusilcpanel_2020 (incl. eusilcpanel_2020.ado; eusilcpanel.sthlp; totalpopulation_2003_2021.dta) is an update of eusilcpanel. It covers the EU-SILC longitudinal releases 2005 to 2020. Both, eusilcpanel and eusilcpanel_2020 can be downloaded here:
https://www.gesis.org/en/missy/materials/EU-SILC/tools/datahandling or here
https://www.gesis.org/gml/european-microdata/eu-silc => eusilcpanel – UPDATE (2022). The original script was based on CSV data. However, since 2017 the CSV files are released by country and year. The CSV data are sometimes subject to noise, therefore eusilcpanel_2020 is based on ‘cleaned’ Stata system files (based on Stata scripts provided by GESIS
https://www.gesis.org/en/missy/materials/EU-SILC/setups).

Please note, as from 2021, the EU-SILC panel design was extended on a voluntary basis to a six-year rotational panel design. However, at the time of updating this paper (2022) the longitudinal Scientific-Use-Files still cover only up to 4 years.

1 Introduction

Each year, more than 500.000 surveyed individuals make EU-SILC one of the world's biggest data collection efforts. Around thirty countries are involved directly in gathering and elaborating observations which leads to a challenging level of complexity. Eurostat releases the results in separate chunks of data covering up to four years (as of 2022). Assembling datasets that cover longer periods of time by hand is a tedious task. There are not many examples of this approach, such as the one made by Engel and Schaffner (2012). Today, many publications use only small portions of the EU-SILC longitudinal dataset, and it remains under-appreciated with respect to similar sources (Eiffe and Till, 2014).

What follows is an attempt to build a tool that is able to provide a single longitudinal dataset (separately for the D-, H-,R-, and P-File) that includes all observations ever collected by EU-SILC. It has the aim of harnessing the potential of more than 9 million comparable observations collected in around 30 European countries from 2003 to 2020. To do so, we use a Stata script that automates the process.

The following paragraphs first highlight some characteristics of the EU-SILC panel that are crucial to building a single dataset. Second, we discuss how datasets from each release can be merged from a theoretical point of view. Then, we suggest how weights should be adjusted. Finally, we describe a Stata script based on these ideas and look at how it performs. We also provide some examples that facilitate the use of the script in practice.[1]

[1] Download: https://www.gesis.org/gml/european-microdata/eu-silc/ (EU-SILC Tools) or: https://www.gesis.org/en/missy/materials/EU-SILC/tools/datahandling.

1 Essential practical information about the EU-SILC datasets

The longitudinal data for EU-SILC is collected following an "integrated" or "rotational" design (p. 16 Eurostat 2015). This means that each country's sample consists of four sub-samples. Each of those sub-samples is observed for four years before it is dropped and a new sub-sample takes its place. In particular, each year one sub-sample leaves the sampling while another one is added (see Figure 1). The reason behind this choice is that the integrated design minimizes practical issues (referred to as "friction") linked to extended periods of following the same households, such as dropouts. Starting with 2021, the panel design was extended on a voluntary basis to a six-year rotational panel design (European Parliament and European Council, 2019, p. 30). The system outlined below remains the same, except that two additional waves are added.

Each year, the EU Member States send Eurostat a file containing only the most recent observations plus past observations of the sub-samples ("rotational groups"), that are still "active". Looking at Figure 1, that would be the observations contained in the grey boxes from *T* to *T*-3. The data contained in box 1 is published as part of the cross-sectional dataset instead, and therefore is not contained in the longitudinal one.

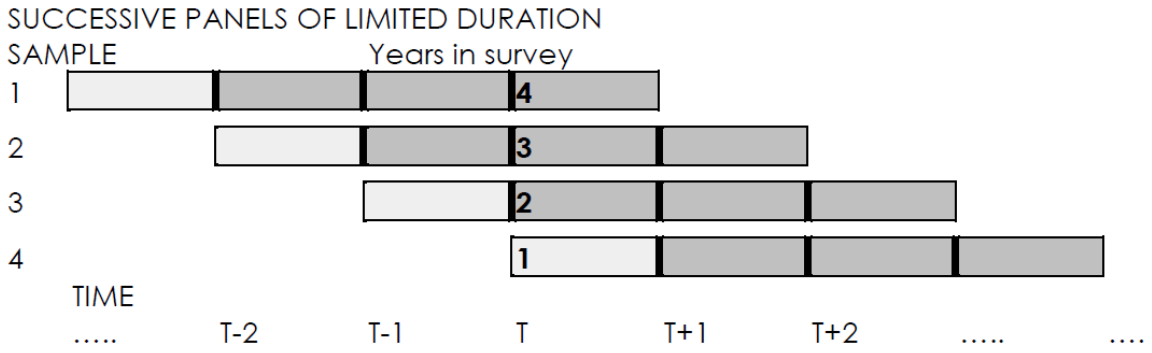


Figure 1: The rotational design. Source: Eurostat, 2015

Each year, the EU-SILC dataset is distributed in four comma separated values (.csv) files, for example for the year 2013:[2]

1. UDB_l13**D**_ver 2013-2 from 01-01-2016.csv
2. UDB_l13**H**_ver 2013-2 from 01-01-2016.csv
3. UDB_l13**R**_ver 2013-2 from 01-01-2016.csv
4. UDB_l13**P**_ver 2013-2 from 01-01-2016.csv

[2] As from the 2016, the csv files are released separately by country, which makes data preparation very time-consuming. However, this does not change the data structure as described in this paper.

They differ from each other by the letters *D*, *H*, *P*, and *R*:

1. The *D* file is a household register, i.e. it contains data about households that were known before the survey, such as household ID, country, region, year of survey, and so on.
2. The *H* file contains all data that has been collected on a household level during the surveys, such as total gross household income and housing costs. There are some households that are present in the register *D*, but then didn't participate in the surveys, so they are missing in the *H* file.
3. The *R* file contains data from the members of the households in *H*. This data has been collected on a household basis as well. It contains similar values with respect to the *D* file, but has more variables.
4. The *P* file contains data that has been collected on an individual level: some individuals are invited to participate in a second survey to collect more data.

Within one release, household and individual datasets are linked between each other through identification numbers that indicate the household an individual is part of. The same IDs can also be used to merge data from the register files with the collected data. So, a household identified by the household ID (*hid*) in *D* can be found again in the *H* file. In the *R* file, we may find the individuals being part of that household, i.e. the individuals that share the same *hid*. In the *P* file are individuals from the *R* file (*Figure 2*).

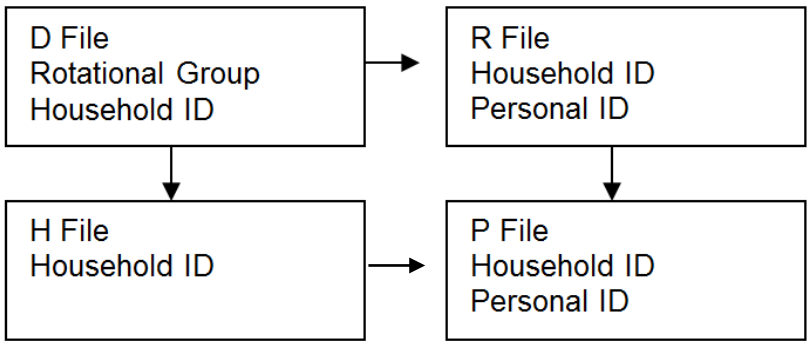


Figure 2: links between the datasets contained in each release

Information regarding rotational groups is contained only in the *D* file. This means, that if an analyst wants to select certain data contained in the *P* file pertaining to a specific rotational group, he must merge the *D*, *R* and *P* file to gather all the necessary information in terms of IDs.

As far as numbers of observations regards, not all households listed in the register file (*D*) file, are also listed in the *H* file and their members might not be in the *R* file. Furthermore, not all individuals contained in the *R* file are also in the *P* file.[3]

The household ID (*DB030* in the *D* file, *hid* in the script) is unique for each household within a release in a given country and year in the *D*- and in the *H* file. The *R* file contains data of individuals, so the *hid* doesn't identify a single observation in a given year and country because each house-

[3] For more on this, see Wirth/Pforr (2022, p. 7).

hold typically comprises several individuals. Neither the personal ID (*pid*) does uniquely identify a single observation in the *R* file because a single individual can be part of more than one household. Only the combination of *country*, *year*, *hid* and *pid* uniquely identifies entries in *R*. In the *P* file the personal ID *pid* does uniquely identify individuals when combined with *country* and *year*. For more on this, see (p.84 Eurostat 2015, "Identification numbers and records of persons").

Hence, a reasonable approach is to select rotational groups according to some criteria (defined later) from the *D* file. Then, one merges the result with the *H* and *R* file to select the data corresponding to those rotational groups. Finally, one merges *R* with *P*.

Table 1: files, IDs and uniqueness

File	Identification numbers contained in file	Unique identification of observations
D file	rotation_group, hid	year & country & hid
H file	hid	year & country & hid
R file	hid, pid	year & country & hid & pid
P file	hid, pid	year & country & pid

Please note: The updated script **eusilcpanel_2020** is based on EU-SILC Stata systemfiles, while the original version **eusilcpanel** is based on the CSV data. In our experience, problems sometimes occur when merging the separate country CSV data in an all-countries-file. The transformation of the CSV files into Stata files was done with the latest Stata scripts (setups) provided by GESIS, German Microdata Lab (https://www.gesis.org/en/missy/materials/EU-SILC/setups).

2 How to build the cumulative dataset: theory

2.1 Selection of rotational groups explained through examples

If a country has been taking part in EU-SILC from 2009, its 2014 release should have the following structure in terms of rotation groups (*Figure 3*).

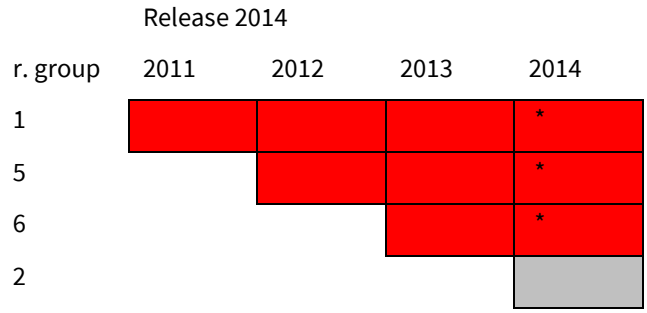
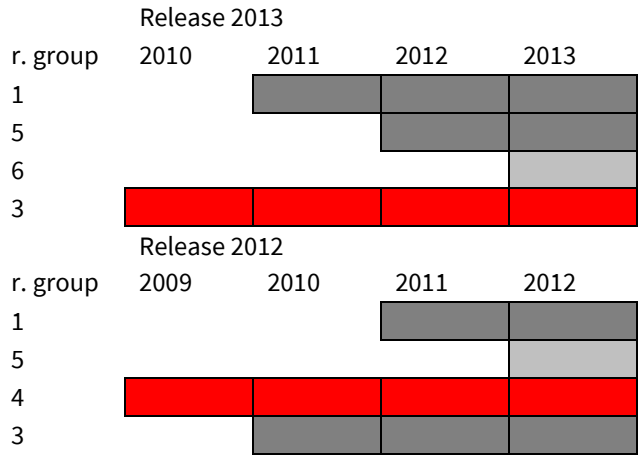


Figure 3: release structure

Group number 2 is not contained in the dataset (light grey box) because it is not a longitudinal sample[4]. To add further data from less recent releases, we want to go to the 2013 release and select the rotation group which was completed in 2013 (group 3), and which would be replaced by the new group 2 in 2014. The selection is done by checking which rotation group contains households with most observations over the years. Next, one would open the release 2012 and get the data from group 4, and so on, moving back in time.



[4] This will eventually lead to the fact that after having merged data from all years, there will be a drop in observations in the last year. In theory, it should be possible to retrieve the data from the cross-sectional dataset, but in practice it turns out that in most of the countries it is not possible selecting group number 2 from the cross-section and be sure that no observations from other groups are included.

To select the right sub-sample it is best to check which sub samples contain households with most observations, as well as making sure that no rotation group is counted twice by verifying which rotation group has been selected from the more recent release. Continuing the example from above, if a country started participating in EU-SILC in 2009, the selection process would be as follows: the maximum number of observations of households would be 3 in the 2011 release, but only group 7 is selected because group 4 has already been added to the full dataset with release 2012. In the 2010 release, the maximum number of observations is 2, but one may select only group 9, since the others have already been selected before.

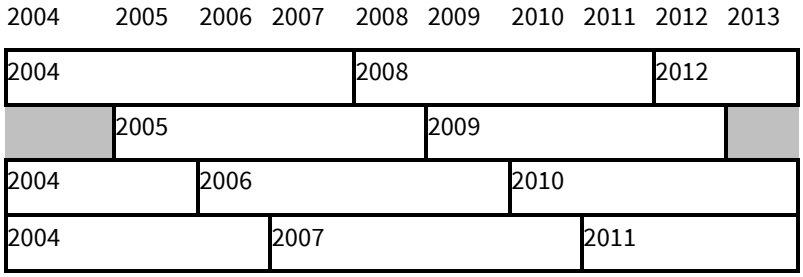
This last selection process is based on comparing the identifier of the rotational group between different releases, which is preferable to the household ID because sometimes observations that had already been collected in earlier releases are dropped in the current release. This is probably due to quality issues in most cases, which is why we assume that the more recent release "overrules" more distant ones. Looking at the example, group 7 might contain 1500 observations for 2008 in the 2010 release, but upon opening the 2009 release we realize that now there are 1800 observations in the exact same cell (group seven, 2008). Thus, there are 300 observations in group 7 that haven't been selected yet, but we still do not want to select them because there is probably a good reason they haven't been reported in 2010.

2.2 Scaling of weights

The EU-SILC panel contains a series of weights that should make sure samples are representative. The documentation (Eurostat, 2015) and Verma (2006) discuss how these weights are calculated. For this work two weights are of interest:

1. *RB060*, the so-called "modified base-weight". Each observation in the EU-SILC data set (*R* file) comes with this weight. In the first year of observation it equals the design weight, calibrated and modified to take non-responses into account. The base-weight of the years that follow is given by the previous year's base-weight adjusted for non-response rates.
2. *RB064* are longitudinal weights that have been created to be used with datasets made up by one rotational group covering four years (within a given release). They are built with the intent of making sure that this sub-sample is representative of the longitudinal population of the year in which the rotational group had been surveyed for the first time. *RB064* is reported only during the last year of a rotational group covering four years. It is constant with respect to the year of observation, but varies across individuals. Since both weights are built based on single rotational groups, they can be used to calculate weights for a larger, cumulative sample. *RB060* can be simply rescaled. The same goes for *RB064*, but in this case, one must restrict the merged sample to rotational groups that cover 4 years, which leads to loss of data. Also, *RB064* makes sure that the sample is representative with respect to the longitudinal population of the year in which the rotational group was first surveyed. This means that the final result becomes something resembling a "moving sample", a set of sub-samples representative of different longitudinal populations. Whether this is desirable or can be avoided by building more sophisticated weights is up for debate. In practice, even though *RB064* and *RB060* take on very different values in some cases, on average the difference is not much.

Table 2: Structure of the full sample for a country 2004-2013. The years in the boxes refer to the longitudinal population that is represented by the rotational group.



The scaling for *RB060* would consist of multiplying the weight by 1/4, provided there were always four rotational groups for each year in the sample and the starting number of observations was always the same. In reality, this is not the case. Hence, a more flexible approach is to use

$$\forall \text{country} \; \forall \text{ year} \quad rscale = \frac{\sum_{rotation\ group} RB060}{\sum_{country} RB060}; \qquad RB060s = rscale \cdot RB060;$$

where *RB060s* is the rescaled weight. *rscale* turns out to be close to ¼ in most cases, except for the "fringes" at the extremes of the cumulative sample where the number of rotational groups drops.

In some countries, the base weight in the *R* file can't be applied directly to data in the *P* file. In these cases, Member States provide a second weight, *PB*080 that can be rescaled in the same way.

The same line of reasoning applies to *RB064* with a minor tweak: as mentioned earlier, *RB064* is reported only in the last year of panel covering four years. So, the first step is to copy *RB064* to all years of all rotational groups that come with *RB064*. Second, analysts should drop all observations with *RB064s* missing to make sure that only rotational groups covering four years are in the sample.

RB064 can't be used with the data from France: the country reports each release six rotational groups covering four years (instead of one year). Each rotational group stays for nine years (instead of four) in the sample, even though each release covers only four years. As a consequence, *RB064* does not make sure that single rotational groups are representative (in fact, four groups jointly form a representative sample). At the same time, one can't just select all four groups in each release because this would lead to overlapping across different releases due to the rotational groups staying for nine years in the sample. For similar reasons (recent changes in the number of rotational groups respectively years of stay in the sample) RB064 should not be used e.g. with data from Belgium and Bulgaria.

3 How to build a cumulative sample: practice

3.1 First: a word of caution

The initial motivation for this project was to build a Stata Package that would be able to merge all releases in an automated fashion and readily deliver a full panel within hours. Unfortunately, not all countries follow exactly the integrated design as suggested by Eurostat. The Stata script here is the result of a process of trial- and error, i.e. of checking for inconsistencies in the results and then applying fixes to the script.

In the case of Iceland (IS), Norway (NO), and also Luxemburg (LU), an automated approach doesn't work despite of fixes because these countries change their sample design across different releases. Other countries, such as France (FR) and Ireland (IE), come with some issues but they could be overcome by applying minor changes to the script. A (probably non-exhaustive) list of issues we found are provided in the Annex. Finally, for Germany panel data are released only recently (1st wave 2015) for scientific purposes.

3.2 System requirements and input preparation

The script has been tested on a system with an Intel i7-2760QM (quad core) processor, 8GB of RAM, a solid state drive (SSD) with at least 30GB of free space, and running StataMP 13 on Windows 10 Pro. Another operation system might lead to problems with regard to how the script automatically finds and opens files. Lower hardware specifications (except for the mandatory free space on the hard disk) may slow down the execution of the script. The base version of Stata is not sufficient because the merging of the *R* and *P* files goes beyond its pre-imposed limits of working memory ("op. sys. refuses to provide memory" error).

As of October 2017, Eurostat implemented changes to the delivery of EU-SILC data. CSV-files are now released by country and year. In order to be able to run the updated script, it is necessary (1) to transform the csv files into Stata-Systemfiles using the setups prepared by GESIS (https://www.gesis.org/en/missy/materials/EU-SILC/setups), (2) to prepare a folder named "EU-SILC" that contains all the Stata-System files from the releases of 2005 until 2020. The directory-tree below represents the file structure the final EU-SILC folder should have. The script is indifferent about the part of filenames of the .dta files following "dta", indicated here by "…".

```
C:\...\EU-SILC
├───1 L-2005
│       UDB_L05D_ver….dta
│       UDB_L05H_ver….dta
│       UDB_L05P_ver….dta
│       UDB_L05R_ver….dta
├───2 L-2006
│       UDB_L06D_ver….dta
│       UDB_L06H_ver….dta
│       UDB_L06P_ver….dta
│       UDB_L06R_ver….dta
├───3 L-2007
│       UDB_l07D_ver….dta
│       UDB_l07H_ver….dta
│       UDB_l07P_ver….dta
│       UDB_l07R_ver….dta
├───4 L-2008
│       UDB_l08D_ver….dta
│       UDB_l08H_ver….dta
│       UDB_l08P_ver….dta
│       UDB_l08R_ver….dta
├───5 L-2009
│       UDB_l09D_ver….dta
│       UDB_l09H_ver….dta
│       UDB_l09P_ver….dta
│       UDB_l09R_ver….dta
├───6 L-2010
│       UDB_l10D_ver….dta
│       UDB_l10H_ver….dta
│       UDB_l10P_ver….dta
│       UDB_l10R_ver….dta
├───7 L-2011
│       UDB_l11D_ver….dta
│       UDB_l11H_ver….dta
│       UDB_l11P_ver….dta
│       UDB_l11R_ver….dta
├───8 L-2012
│       UDB_l12D_ver….dta
│       UDB_l12H_ver….dta
│       UDB_l12P_ver….dta
│       UDB_l12R_ver….dta
├─── (...)
│       (...)
│       (...)
│       (...)
│       (...)
├───15 L-2015
│       UDB_l15D_ver….dta
│       UDB_l15H_ver….dta
│       UDB_l15P_ver….dta
│       UDB_l15R_ver….dta
└───16 L-2020
        UDB_l20D_ver….dta
        UDB_l20H_ver….dta
        UDB_l20P_ver….dta
        UDB_l20R_ver….dta
```

Once the EU-SILC folder is ready, one can run the script by saving the files (eusilcpanel_2020.ado, eusilcpanel.sthlp, totalpopulation_2003_2021.dta) in …/ado/personal, typing `eusilcpanel_2020` in the Stata command window, and inserting the path of the EU-SILC folder, as prompted.

Depending on the system specifications, it will take several hours before the final result is available. The script contains a number of non-critical checks that are handy for troubleshooting, even though they have a negative impact on computational efficiency.

3.3 Some notions about how "eusilcpanel" works

This paragraph gives a very general description of how the script works. For further details, refer to the comments in the code and the annex.

The script starts by loading the 2020 *D* file; it selects all data from this release, and then generates a series of new variables (see 4.4 Output) that are needed further down the road.

Next, it opens the *D* file from the 2019 release and selects those rotation groups that cover most years. It also checks whether a given group has already been selected from the more recent (in this case 2020) release, i.e. whether a given year and rotational group has already been covered. If yes, the rotational group is not selected in the years in question (as happens often in case of France). If not, the rotational group is selected. This is to make sure all data is captured in case of countries that took part in the 2019 release, but not in the 2020 release. Now, the D file in memory (2019) is merged with the more recent *D* file (2020). One final check controls for cases in which the ID of a rotational group has changed across different releases by checking for duplicates in terms of household IDs.

The process above is repeated for all releases, with the only difference that as we move further back in time, the checks described above need to be performed not only with respect to the most recent release, but with respect to two or more recent releases. The result of this process is the masterD.dta file.

Next, the script chooses from the *H* files households that are in the masterD.dta file. The same applies to the *R* file. Observations in the *P* file are selected based on the content of the masterR.dta file. Finally, the script rescales the weights in *R*-, and then in the *P* file.

3.4 Output

eusilcpanel_2020 produces four files which are saved in the "16 L-2020" folder: masterD.dta, masterH.dta, masterR.dta, and masterP.dta. They contain the same variables as the respective UDB files. On top of that, they contain some additional values:

year = db010, hb010, rb010, pb010 (year of observation);

country = db020, hb020, rb020, pb020 (country of observation);

hid = *db030, hb030, rb040* (household ID) provided by Eurostat/Member States;

rotation_group = *db075*, rotation group ID provided by Eurostat/Member States;

urtgrp is an alternative ID for rotational groups that is unique across all countries and releases. It is a string composed of *country*, *rotation_group* and the last year in which the rotation group was (or will be) active assuming 4 years of observation;

uhid is an alternative household ID that is unique across all releases and countries. It is a string composed of *urtgrp* and the household ID hid (db030);

uhidnum is a numerical household ID based on *uhid*. Being numerical it can be used with `xtset`;

pid = *rb030* personal ID provided by Eurostat/Member States;

upid is a unique personal ID composed by the first 7 places of the unique household ID (*uhid*) and *pid*. Note that *pid* does not uniquely define units in the *R* file because one individual can be part of more than one household at the same time. This means that there are individuals with more than one *upid* in the *R* file;

upidnum is a numerical personal ID based on *upid*. Being numerical it can be used with `xtset`.

pop is the population size of a country during a given year, sourced from Eurostat[5];

rscale is the sum of base weights (rb060) of a rotational group divided by the sum of base weights of a country;

rb060s: base weight (*rb060*) multiplied by *rscale*. These weights can be used as base weights in the merged dataset.

smwrate60: is the difference between sum of weights (*rb060s*) and total population size (*pop*) divided by total population size.

lrscale: is the sum of longitudinal weights (rb064) of a rotational group divided by the sum of longitudinal weights of a country;

rb064s: longitudinal weight (*rb064*) multiplied by *lrscale*. These weights can be used as longitudinal weights in a merged dataset.

smwrate64 is the difference between sum of weights (*rb064s*) and total population size (*pop*) divided by total population size.

pscale: sum of individual weights (*pb080*) of a rotational group divided by the sum of weights of a country;

pb080s: personal weight (*pb080*) multiplied by *pscale*;

smwrate80: is the difference between sum of weights (*pb80s*) and total population size (*pop*) divided by total population size.

[5] http://ec.europa.eu/eurostat/tgm/table.do?tab=table&language=en&pcode=tps00001&tableSelection=1&footnotes=yes&labeling=labels&plugin=1

3.4.1 Output Analysis

The script is not able to provide a dataset one can use as-is. It provides a set of observations from which analysts may choose subsets that seem to be good samples. What follows are a series of Stata commands that may help to find a good subset.

`tab country year` produces a table that shows the number of observations of each year and country. It shows how different countries started collecting data at different points of time and is a good starting point for any analysis.

	year																	
country	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
AT	0	11,550	13,351	15,113	17,035	13,935	13,882	14,373	14,184	14,193	13,519	13,182	13,433	13,250	13,107	12,984	12,527	8,283	227,901
BE	0	9,902	12,820	13,970	15,018	14,590	14,514	14,716	14,558	14,334	15,001	14,564	14,448	14,056	14,266	13,953	15,877	13,843	240,430
BG	0	0	0	9,091	11,631	11,893	14,403	15,848	17,021	14,994	16,009	15,491	15,133	18,155	18,089	17,441	17,439	13,726	226,364
CH	0	0	0	0	0	0	0	0	5,023	9,129	12,866	15,943	17,444	18,232	19,085	15,517	16,982	11,546	141,767
CY	0	0	8,506	11,322	10,836	10,199	9,461	11,283	11,674	13,686	13,572	12,235	12,201	11,442	11,700	11,269	11,146	8,065	178,597
CZ	0	0	10,333	18,050	23,313	27,165	23,565	21,648	20,893	20,484	19,345	18,337	17,872	19,314	19,743	19,486	19,630	14,210	313,388
DE	0	0	0	0	0	0	0	0	0	0	0	0	8,551	15,554	21,371	25,355	16,338	0	87,169
DK	4,092	7,099	9,983	12,192	10,712	9,816	9,284	8,931	8,308	8,635	9,607	9,871	9,723	8,937	8,210	8,319	10,124	8,158	162,001
EE	0	11,665	12,222	16,272	14,703	13,306	13,935	13,810	13,713	14,582	15,384	15,420	14,945	15,428	15,672	15,315	15,511	11,135	243,018
EL	0	0	0	4,739	8,463	14,072	18,323	17,913	15,463	13,768	18,039	21,166	34,709	44,623	54,782	57,755	40,520	26,992	391,327
ES	0	33,851	38,527	35,526	35,460	36,858	37,644	37,930	35,629	34,355	32,867	32,285	33,090	32,952	31,335	30,201	36,078	22,396	576,984
FI	0	15,474	19,741	18,499	17,639	17,062	16,266	19,755	23,627	26,019	28,717	27,910	27,135	26,703	25,494	24,536	23,789	16,128	374,494
FR	0	35,838	37,088	25,453	26,421	25,979	26,144	26,983	33,313	36,948	32,984	29,257	27,182	27,168	25,931	25,230	26,977	12,388	481,284
HR	0	0	0	0	0	0	0	9,862	12,617	11,792	14,179	14,237	17,420	19,919	20,492	21,694	19,925	13,763	175,900
HU	0	0	14,567	20,283	22,582	22,685	24,732	24,887	29,729	28,888	26,064	23,225	19,092	19,259	19,021	17,239	15,596	9,896	337,745
IE	0	7,473	10,840	10,516	10,060	11,948	10,852	7,146	8,791	10,088	10,049	12,946	13,385	13,142	9,118	7,888	10,299	6,697	171,238
IS	0	5,944	8,060	7,370	7,138	7,103	7,142	7,590	12,102	14,686	12,631	17,268	12,589	8,323	8,745	6,637	0	0	143,328
IT	0	46,589	57,077	55,357	53,534	53,272	51,902	48,010	49,070	48,410	45,200	47,802	43,831	48,943	49,338	46,552	31,241	0	776,128
LT	0	0	9,100	12,448	13,072	12,379	13,136	13,468	12,879	13,005	12,023	12,118	11,306	11,157	11,395	11,396	11,665	8,411	188,958
LU	19,160	19,696	19,744	20,712	20,794	20,362	20,069	13,552	15,135	16,394	10,173	10,066	8,852	10,292	10,887	10,710	10,702	7,690	264,990
LV	0	0	9,018	11,260	11,468	13,451	14,807	15,856	16,363	15,723	15,107	14,424	14,292	14,198	13,882	13,184	8,703	9,995	211,731
MT	0	0	0	3,376	6,219	8,027	10,309	10,533	11,347	12,103	12,138	12,012	11,430	10,961	10,312	9,963	9,712	6,789	145,231
NL	0	0	21,634	23,371	26,202	25,739	23,973	24,916	25,758	25,255	25,019	24,844	23,740	29,871	30,080	27,999	30,307	19,351	408,059
NO	11,427	14,005	15,722	14,524	14,710	13,988	13,562	12,018	13,786	22,872	22,141	22,123	15,639	16,801	14,845	14,271	14,638	10,055	277,127
PL	0	0	36,525	46,022	43,629	42,071	39,365	38,056	37,448	37,861	37,153	36,786	34,337	33,188	35,377	40,587	51,490	40,037	629,932
PT	0	7,092	9,921	12,251	11,857	11,985	13,175	13,576	14,860	16,229	16,659	17,469	22,252	26,929	30,423	34,429	33,600	18,528	311,235
RO	0	0	0	0	14,902	19,219	18,844	18,422	18,047	17,777	17,707	17,387	17,419	17,382	17,234	17,158	16,819	12,693	241,010
RS	0	0	0	0	0	0	0	0	0	0	15,175	19,345	18,544	18,005	16,918	16,765	16,380	11,596	132,728
SE	0	13,734	18,992	16,240	16,358	16,205	16,072	15,688	14,150	14,665	14,271	14,280	14,447	14,217	14,821	14,545	13,685	9,937	252,307
SI	0	0	27,670	31,899	28,885	29,492	30,179	30,127	29,337	28,634	27,786	28,190	26,595	26,112	26,758	26,269	25,693	16,501	440,127
SK	0	0	11,613	14,972	14,188	15,225	15,865	15,760	15,454	15,539	15,569	15,734	16,303	16,649	12,301	15,918	10,786	10,104	231,980
UK	0	0	29,880	34,689	22,278	21,463	19,720	19,060	18,948	26,337	26,196	27,667	27,128	27,961	38,381	16,840	0	0	356,548
Total	34,679	239,912	462,934	515,517	529,107	539,489	541,125	541,717	569,227	597,385	603,150	613,584	614,467	649,123	669,113	647,405	594,179	378,913	9,341,026

Figure 4: tab country year, Stata page print, masterR.dta

Next, one might zoom in on specific countries. `tab urtgrp year if country == "AT"` for example, shows which rotational groups cover which years in the case of Austria.

	year																
urtgrp	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
AT12008	0	4,818	4,225	3,827	2,785	0	0	0	0	0	0	0	0	0	0	0	0	15,655
AT12012	0	0	0	0	0	4,561	3,977	3,540	3,200	0	0	0	0	0	0	0	0	15,278
AT12016	0	0	0	0	0	0	0	0	0	4,079	3,259	2,963	2,694	0	0	0	0	12,995
AT12020	0	0	0	0	0	0	0	0	0	0	0	0	0	3,928	3,239	2,760	2,599	12,526
AT22007	2,824	2,155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4,979
AT22009	0	0	4,964	4,301	3,135	2,738	0	0	0	0	0	0	0	0	0	0	0	15,138
AT22013	0	0	0	0	0	0	4,450	3,817	3,480	3,090	0	0	0	0	0	0	0	14,837
AT22017	0	0	0	0	0	0	0	0	0	0	4,247	3,633	3,109	2,891	0	0	0	13,880
AT22021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4,009	3,101	2,665	9,775
AT32007	2,910	2,166	1,998	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7,074
AT32010	0	0	0	5,169	3,737	3,147	2,904	0	0	0	0	0	0	0	0	0	0	14,957
AT32014	0	0	0	0	0	0	0	4,080	3,457	3,044	2,739	0	0	0	0	0	0	13,320
AT32018	0	0	0	0	0	0	0	0	0	0	0	4,047	3,164	2,764	2,607	0	0	12,582
AT32022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3,796	3,019	6,815
AT42007	5,816	4,212	3,926	3,738	0	0	0	0	0	0	0	0	0	0	0	0	0	17,692
AT42011	0	0	0	0	4,278	3,436	3,042	2,747	0	0	0	0	0	0	0	0	0	13,503
AT42015	0	0	0	0	0	0	0	0	4,056	3,306	2,937	2,790	0	0	0	0	0	13,089
AT42019	0	0	0	0	0	0	0	0	0	0	0	0	4,283	3,524	3,129	2,870	0	13,806
Total	11,550	13,351	15,113	17,035	13,935	13,882	14,373	14,184	14,193	13,519	13,182	13,433	13,250	13,107	12,984	12,527	8,283	227,901

Figure 5: tab urtgrp year if country == "AT", Stata page print, masterR.dta

The resemblance of Figure 5 with *Table 2* is an indication that the script has done a good job. One also notices that at the extremes there are less rotational groups covering a given year. Even though the rescaling of the weights takes this into account, one still might prefer to choose only years 2005 to 2019 to obtain a smaller, but potentially more representative sample.

`tab urtgrp year if country == "IE"` shows the rotational groups of Ireland. In this case, one can see some irregularities: 2010 contains only two rotational groups, while 2011 contains three.

	year																
urtgrp	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
IE12007	1,952	2,856	3,408	2,641	0	0	0	0	0	0	0	0	0	0	0	0	0	10,857
IE12011	0	0	0	0	2,844	1,945	0	0	0	0	0	0	0	0	0	0	0	4,789
IE12015	0	0	0	0	0	0	0	0	3,841	3,249	1,626	926	0	0	0	0	0	9,642
IE12020	0	0	0	0	0	0	0	0	0	0	0	0	3,458	1,357	1,351	1,337	787	8,290
IE22007	2,945	3,819	2,772	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9,536
IE22010	0	0	0	2,162	3,266	2,171	0	0	0	0	0	0	0	0	0	0	0	7,599
IE22014	0	0	0	0	0	0	0	3,584	3,112	2,046	1,045	0	0	0	0	0	0	9,787
IE22018	0	0	0	0	0	0	0	0	0	0	0	4,041	3,286	2,561	1,805	0	0	11,693
IE22022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4,189	2,516	6,705
IE32007	2,576	2,265	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4,841
IE32009	0	0	1,546	2,436	3,407	2,294	0	0	0	0	0	0	0	0	0	0	0	9,683
IE32013	0	0	0	0	0	0	3,629	3,205	2,076	1,095	0	0	0	0	0	0	0	10,005
IE32017	0	0	0	0	0	0	0	0	0	0	7,124	5,923	4,519	3,610	0	0	0	21,176
IE32021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3,113	3,172	2,229	8,514
IE42008	0	1,900	2,790	2,821	2,431	0	0	0	0	0	0	0	0	0	0	0	0	9,942
IE42012	0	0	0	0	0	4,442	3,517	2,002	1,059	0	0	0	0	0	0	0	0	11,020
IE42016	0	0	0	0	0	0	0	0	0	3,659	3,151	2,495	1,879	0	0	0	0	11,184
IE42020	0	0	0	0	0	0	0	0	0	0	0	0	0	1,590	1,619	1,601	1,165	5,975
Total	7,473	10,840	10,516	10,060	11,948	10,852	7,146	8,791	10,088	10,049	12,946	13,385	13,142	9,118	7,888	10,299	6,697	171,238

Figure 6: tab urtgrp year if country == "IE", Stata page print, masterR.dta

What happened? By using masterD.dta and running `tab urtgrp yrelease if country == "IE"`, it turns out that no data has been collected from Ireland in the 2010 and 2011 release because Ireland didn't contribute to these releases.

	yrelease													
urtgrp	2005	2006	2007	2008	2009	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
IE12007	0	0	4,695	0	0	0	0	0	0	0	0	0	0	0	4,695
IE12011	0	0	0	0	2,161	0	0	0	0	0	0	0	0	0	2,161
IE12015	0	0	0	0	0	0	0	0	4,783	0	0	0	0	0	4,783
IE12020	0	0	0	0	0	0	0	0	0	0	0	0	1,318	2,172	3,490
IE22007	0	4,139	0	0	0	0	0	0	0	0	0	0	0	0	4,139
IE22010	0	0	0	0	3,306	0	0	0	0	0	0	0	0	0	3,306
IE22014	0	0	0	0	0	0	0	4,703	0	0	0	0	0	0	4,703
IE22018	0	0	0	0	0	0	0	0	0	0	0	5,536	0	0	5,536
IE22022	0	0	0	0	0	0	0	0	0	0	0	0	0	3,220	3,220
IE32007	1,958	0	0	0	0	0	0	0	0	0	0	0	0	0	1,958
IE32009	0	0	0	0	4,086	0	0	0	0	0	0	0	0	0	4,086
IE32013	0	0	0	0	0	0	4,874	0	0	0	0	0	0	0	4,874
IE32017	0	0	0	0	0	0	0	0	0	0	9,519	0	0	0	9,519
IE32021	0	0	0	0	0	0	0	0	0	0	0	0	0	3,636	3,636
IE42008	0	0	0	4,419	0	0	0	0	0	0	0	0	0	0	4,419
IE42012	0	0	0	0	0	5,542	0	0	0	0	0	0	0	0	5,542
IE42016	0	0	0	0	0	0	0	0	0	5,081	0	0	0	0	5,081
IE42020	0	0	0	0	0	0	0	0	0	0	0	0	0	2,496	2,496
Total	1,958	4,139	4,695	4,419	9,553	5,542	4,874	4,703	4,783	5,081	9,519	5,536	1,318	11,524	77,644

Figure 7: tab urtgrp yrelease if country == "IE", Stata page print, masterD.dta [Please note: Until 2019, Ireland used a four-year rotating panel. Due to the new regulation, Ireland increased the rotation panel to 5 years. 2020 is the first year to include five waves. As of 2022 the rotation pattern will be increased to six waves: https://www.cso.ie/en/releasesandpublications/in/silc/informationnote-breakintimeseriessilc2020/]

The script limits the damage by selecting those rotational groups from the 2009 release that should have been selected from the 2010 and 2011 releases, but still, some data is missing. Once again, the rescaled weights compensate for drop in observations to some degree: using the masterR.dta and keeping only data from Ireland, `tab year , summarize ( rscale )` reports the average scaling factor for the base weights of a given year. Since in 2011 there are only three rotational groups in the sample and in 2010 there are only two, the scaling factor varies accordingly.

-> country = IE

year	Summary of rscale Mean	Std. dev.	Freq.
2004	.32764619	.23765232	7,473
2005	.26216348	.08940533	10,840
2006	.24704953	.01432036	10,516
2007	.24773481	.02747194	10,060
2008	.25039358	.00419931	11,948
2009	.25052246	.00179383	10,852
2010	.5003504	.02235665	7,146
2011	.33685512	.01494117	8,791
2012	.24856313	.00596175	10,088
2013	.24828305	.00688987	10,049
2014	.25594443	.01921424	12,946
2015	.26754321	.02620915	13,385
2016	.24993059	.00794987	13,142
2017	.3385152	.23394678	9,118
2018	.2288286	.42010528	7,888
Total	.27684353	.13853587	154,242

Figure 8: tab year , summarize (rscale) , Stata page print, masterR.dta (Ireland only); RB060 for IE yet not included in the release EU-SILC 2020_ver_2022_03.

After choosing a sample, one can work with the master-files in the same way in which one would work with the single UDB files. The only difference is that extra care must be placed in looking for issues with single variables across different releases. The “Problems and Modifications” spreadsheets that come with each release provided by Eurostat are a good starting point.

4 Conclusion

The idea at the beginning was to write a Stata package that would automatically produce a cumulative sample analysts can use right away. This proved to be more challenging than expected – it turns out that in some cases countries (in particular Norway, Iceland and Luxembourg) use sample designs that change across releases or are so different from the standard EU-SILC rotational design that an automated approach makes little sense. Even selecting rotational groups from different releases by hand and merging them can lead to samples that are not representative.

Nevertheless, we hope that the script is a rudimentary but helpful tool to explore the full EU-SILC longitudinal dataset. Future efforts should focus on developing better weights. Also, the script has the potential to become computationally much more efficient, perhaps so efficient that it can run on Stata's base version. Finally, analyzing and mapping issues linked to single variables should it be helpful in order to reduce the time spent preparing samples.

Acknowledgements

Carlo D'Ippoliti and Marco Alfò alongside their colleagues from the Department of Statistics at the University of Rome "La Sapienza" provided essential assistance and support during all stages of this project.

Bibliography

Eurostat, 2015, DESCRIPTION OF TARGET VARIABLES: Cross-sectional and Longitudinal 2015 operation (Version August 2016)

Eiffe F. and Till M., 2014, The Longitudinal Component of EU-SILC: Still Underused?, Working Paper 1/2014, NetSILC2

Engel M. and Schaffner S., 2012, How to Use the EU-SILC Panel to Analyze Monthly and Hourly Wages, RUHR Economic Papers

European Parliament and European Council (2019) Regulation (EU) 2019/1700.

Verma V., Betti G, Ghellini G., 2006, Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC Working Paper n. 67, December 2006

Wirth, H. and Pforr, K. , The European Union Statistics on Income and Living Conditions after 15 Years, European Sociological Review, 2022;, jcac024, https://doi.org/10.1093/esr/jcac024

Annex: Issues and Fixes

While writing the script several issues became clear. What follows is a list of problems (and solutions when applicable). The list may be non-exhaustive and new versions of the same releases may potentially lead to new problems or solve some of the problems reported here.

1. Stata has problems with IDs made of large numbers because it rounds them. The solution is to transform IDs from numbers into strings.
2. Sometimes countries lack from more recent releases, but not from earlier ones. This is the case with Croatia (HR) (missing for 2012), Slovakia (SK) (2017, 2019) and IE (2010, 2011). The script recovers some of them when jumping to the less recent release (2013 in HR) by selecting not only the rotational group covering most years, but also groups that haven't been selected before.
3. Some countries (ES, FI, FR, IS, LU, NO, PT, RO, SK, UK, SE) don't follow strictly the rotational design. Simply selecting rotational groups leads to selecting the same observation more than once. Therefore, a further mechanism is needed that checks whether the same observation in the same rotational group (defined by *hid* and *rotation_group*) had already been selected. The solution to this problem depends on the specific rotational design of the country:
    a. FR: France uses "prolonged" rotational groups, i.e. a rotational group stays for 9 years instead of 4 in the sample. Still, each UDB file covers only 4 years. This means that a given rotational group, for example, covers 2012-2015 in the 2015 release. Then, in the 2014 release it covers 2011-2014. Without correction, the script would select all data from both releases as if it was created by two different rotational groups. A similar problem arises for countries that switched from a four-year panel to a 5 or 6-year panel during the observation period, e.g., Bulgaria and Belgium. To overcome this problem, the script checks whether observations with the same year and household ID have already been selected from more recent UDB files and then updates *urtgrp* and *uhid*.

       As a consequence, in France, Bulgaria and Belgium, rotation groups are cut and pieced together across different releases. This makes it impossible to use the longitudinal weight *rb064s* with these countries. Also, *urtgrp* contains the drop-out year which is not accurate in case rotation groups contained in the 2015 release of France.
    b. Croatia (HR), Serbia (RS): there are inconsistencies with the guidelines on how the data should be structured, and the number of observations may vary strongly within the same cell across releases. The script works correctly, but the resulting samples can be unbalanced.
    c. Spain (ES), Finland (FI) and Portugal (PT) present cases in which *hid* is not unique when different releases are merged. The script checks whether this is due to a change in the ID of the rotation group or simply because these countries re-use the household IDs. If there is a change in the ID of the rotation group and the group has already been selected from more recent release, the rotation group is dropped to avoid overlapping.
    d. Sweden (SE) presents three new rotational groups in the 2011 release with respect to previous years: rotational group "8", covering four years (as expected), and rotational group "1" (covering three years) and "2" (two years). The script selects all of them. This may lead to an imbalanced sample.
    e. Luxembourg (LU), Norway (NO) and Iceland (IS) change their sample design across different releases. For now there is no convincing way to make the script work with these countries while assuring at the same time acceptable results, also because these issues are intertwined with others, such as changing IDs of rotational groups across different releases.

4. Greece: the variable country at the beginning is GR, then changes to EL in more recent releases starting from 2008.
5. Luxembourg - Sample selection scheme: The cross-sectional sample for LU-SILC 2018 is composed of three panel samples of individuals, which were selected in 2015, 2016 and 2017, and a simple random sample of individuals aged 18 or more drawn in 2018 from Luxembourg's National Population Register (Registre National des Personnes Physiques - RNPP). - A first longitudinal sample (DB075=2) includes 1151 persons who participated in the survey for the first time in 2015; - A second longitudinal sample (DB075=3) includes 2887 persons who participated in the survey for the first time in 2016; - A second longitudinal sample (DB075=4) includes 2908 persons who participated in the survey for the first time in 2017; - A new sample (DB075=1) is composed of 5 000 individuals aged 18 or more who were selected for the first time in 2018. It must be emphasized that those sub-samples cannot be pooled together as the sampling units are different. The combination of the sub-samples is done during the weighting process, through the use of the Weight Share Method (cf annex in the country specific quality annex on the weighting procedure).