gesis

Leibniz-Institut
für Sozialwissenschaften

GESIS Papers

2018|11

EU-SILC Tools: eusilcpanel

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

Marwin Borst

GESIS Papers 2018|**11**

EU-SILC Tools: eusilcpanel

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

Marwin Borst

GESIS – Leibniz-Institut für Sozialwissenschaften 2018

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 469
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 (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.

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. 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 6 million comparable observations collected in around 30 European countries from 2003 to 2015. 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. The script can be downloaded on https://www.gesis.org/gml/european-microdata/eu-silc/ (EU-SILC Tools).

2 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 subsamples 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.

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.

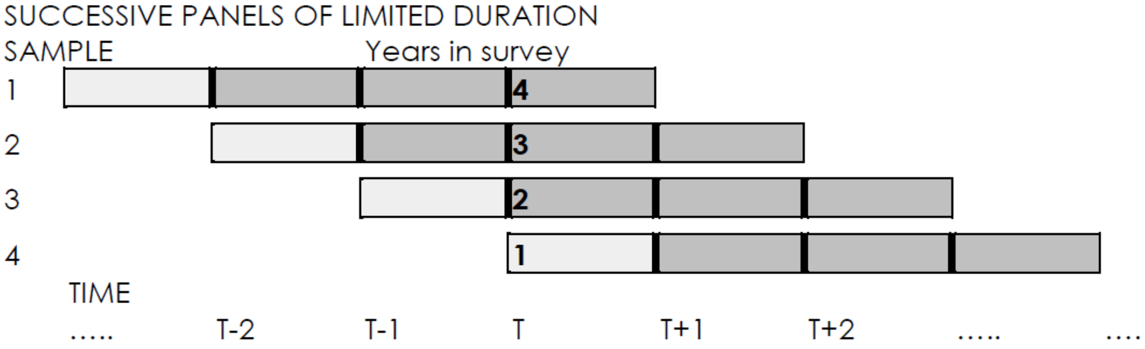


Figure 1: The rotational design. Source: Eurostat, 2015

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

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

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).

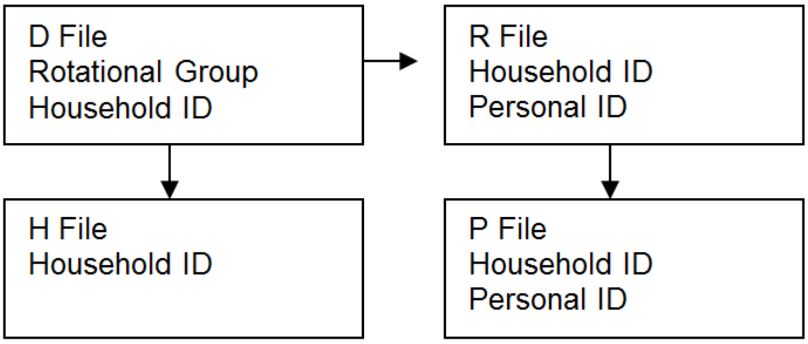


Figure 2: links between the datasets contained in each release

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.

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 household 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

3 How to build the cumulative dataset: theory

3.1 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).

Release 2014

r. group	2011	2012	2013	2014
1	🟥	🟥	🟥	🟥 *
5		🟥	🟥	🟥 *
6			🟥	🟥 *
2				░

Figure 3: release structure

Group number 2 is not contained in the dataset (light grey box) because it is not a longitudinal sample[1]. 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.

Release 2013

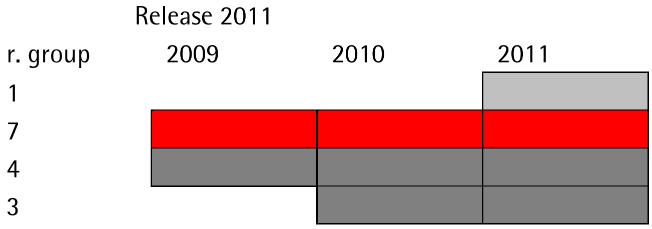
r. group	2010	2011	2012	2013
1		▓	▓	▓
5			▓	▓
6				░
3	🟥	🟥	🟥	🟥

Release 2012

r. group	2009	2010	2011	2012
1			▓	▓
5				░
4	🟥	🟥	🟥	🟥
3		▓	▓	▓

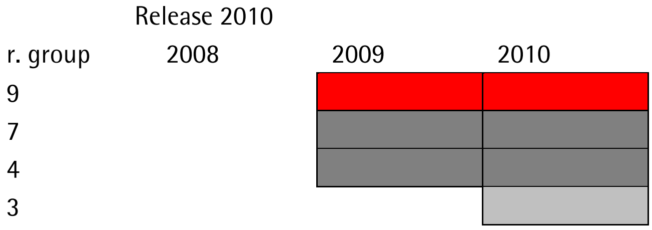
[1] This will eventually lead to the fact that after having merged data from all years, there will be a drop in observations in the last year. In theory, it should be possible to retrieve the data from the cross-sectional dataset, but in practice it turns out that in most of the countries it is not possible 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.



Release 2011

r. group	2009	2010	2011
1			░
7	🟥	🟥	🟥
4	▓	▓	▓
3		▓	▓



Release 2010

r. group	2008	2009	2010
9		🟥	🟥
7		▓	▓
4		▓	▓
3			░

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

3.2 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.

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
2004				2008				2012	
▨	2005				2009				▨
2004		2006				2010			
2004			2007				2011		

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

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

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 eight 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 eight years in the sample.

4 How to build a cumulative sample: practice

4.1 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, we didn't have access to the panel data of Germany, because Germany (up to now) does not release these data (due to data confidentiality reasons).

4.2 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).

In order to be able to run the script, it is necessary to prepare a folder named "EU-SILC" that contains all the decrypted and un-zipped UDB files from the releases of 2005 until 2015. In most cases the preparation consists in simply putting the unzipped "X L-20XX" folders from the releases into the EU-SILC folder and getting rid of any superfluous files such as documentation. Only the folders from the last (2015) release may require some renaming as they contain "minus_DE" and come with no number in the first place. 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 .csv files following "ver", indicated here by "…".

```
C:\...\EU-SILC
├────1 L-2005
│        UDB_L05D_ver….csv
│        UDB_L05H_ver….csv
│        UDB_L05P_ver….csv
│        UDB_L05R_ver….csv
├────2 L-2006
│        UDB_L06D_ver….csv
│        UDB_L06H_ver….csv
│        UDB_L06P_ver….csv
│        UDB_L06R_ver….csv
├────3 L-2007
│        UDB_l07D_ver….csv
│        UDB_l07H_ver….csv
│        UDB_l07P_ver….csv
│        UDB_l07R_ver….csv
├────4 L-2008
│        UDB_l08D_ver….csv
│        UDB_l08H_ver….csv
│        UDB_l08P_ver….csv
│        UDB_l08R_ver….csv
├────5 L-2009
│        UDB_l09D_ver….csv
│        UDB_l09H_ver….csv
│        UDB_l09P_ver….csv
│        UDB_l09R_ver….csv
├────6 L-2010
│        UDB_l10D_ver….csv
│        UDB_l10H_ver….csv
│        UDB_l10P_ver….csv
│        UDB_l10R_ver….csv
├────7 L-2011
│        UDB_l11D_ver….csv
│        UDB_l11H_ver….csv
│        UDB_l11P_ver….csv
│        UDB_l11R_ver….csv
├────8 L-2012
│        UDB_l12D_ver….csv
│        UDB_l12H_ver….csv
│        UDB_l12P_ver….csv
│        UDB_l12R_ver….csv
├────9 L-2013
│        UDB_l13D_ver….csv
│        UDB_l13H_ver….csv
│        UDB_l13P_ver….csv
│        UDB_l13R_ver….csv
├────10 L-2014
│        UDB_l14D_ver….csv
│        UDB_l14H_ver….csv
│        UDB_l14P_ver….csv
│        UDB_l14R_ver….csv
└────11 L-2015
         UDB_l15D_ver….csv
         UDB_l15H_ver….csv
         UDB_l15P_ver….csv
         UDB_l15R_ver….csv
```

Once the EU-SILC folder is ready, one can run the script by saving the files (eusilcpanel.ado, eusilcpanel.sthlp, totalpopulation.dta) in …/ado/personal, typing `eusilcpanel` 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.

4.3 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 2015 *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 2014 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 2015) release, i.e. whether a given year and rotational group has already been covered. If yes, the rotational group is unselected 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 2014 release, but not in the 2015 release. Now, the D file in memory (2014) is merged with the more recent *D* file (2015). 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 more recent release, but with respect to two or more 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.

4.4 Output

eusilcpanel produces four files which are saved in the "11 L-2015" 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[2];

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.

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

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;

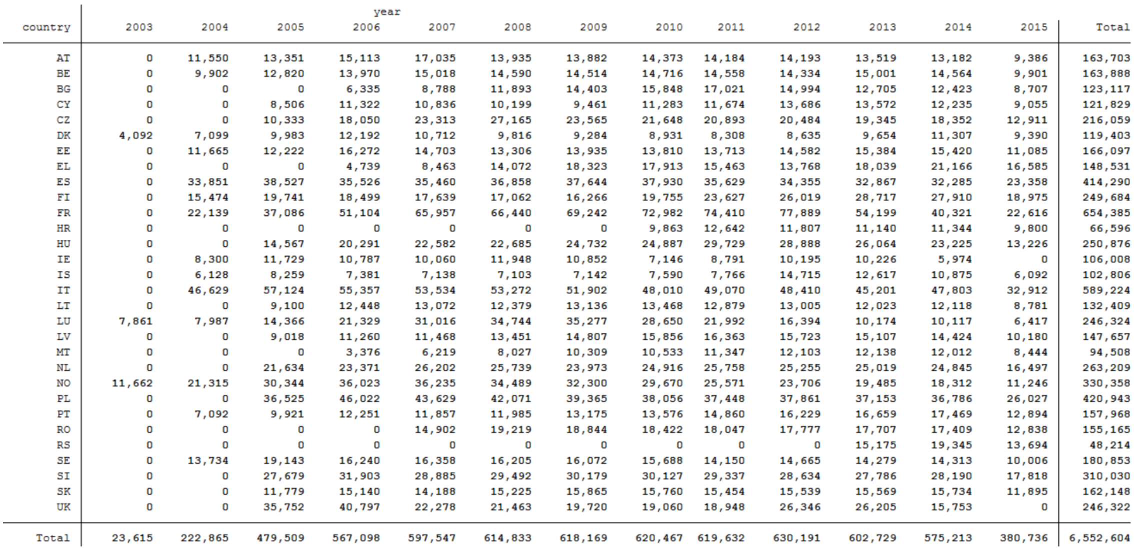
pb080s: personal weight (*pb080*) multiplied by *pscale*;

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

4.4.1 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.

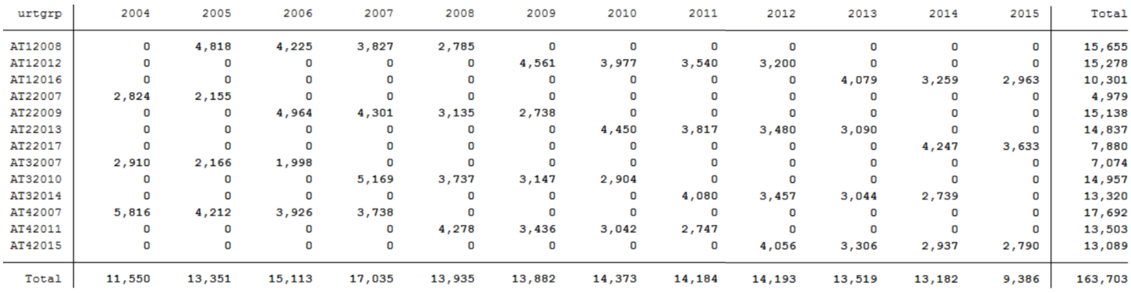
`tab country year` produces a table that shows the number of observations of each year and country. It shows how different countries started collecting data at different points of time and is a good starting point for any analysis.



```
           |                                                 year
   country |      2003      2004      2005      2006      2007      2008      2009      2010      2011      2012      2013      2014      2015 |     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     9,386 |   163,703
        BE |         0     9,902    12,820    13,970    15,018    14,590    14,514    14,716    14,558    14,334    15,001    14,564     9,901 |   163,888
        BG |         0         0         0     6,335     8,788    11,893    14,403    15,848    17,021    14,994    12,705    12,423     8,707 |   123,117
        CY |         0         0     8,506    11,322    10,836    10,199     9,461    11,283    11,674    13,686    13,572    12,235     9,055 |   121,829
        CZ |         0         0    10,333    18,050    23,313    27,165    23,565    21,648    20,893    20,484    19,345    18,352    12,911 |   216,059
        DK |     4,092     7,099     9,983    12,192    10,712     9,816     9,284     8,931     8,308     8,635     9,654    11,307     9,390 |   119,403
        EE |         0    11,665    12,222    16,272    14,703    13,306    13,935    13,810    13,713    14,582    15,384    15,420    11,085 |   166,097
        EL |         0         0         0     4,739     8,463    14,072    18,323    17,913    15,463    13,768    18,039    21,166    16,585 |   148,531
        ES |         0    33,851    38,527    35,526    35,460    36,858    37,644    37,930    35,629    34,355    32,867    32,285    23,358 |   414,290
        FI |         0    15,474    19,741    18,499    17,639    17,062    16,266    19,755    23,627    26,019    28,717    27,910    18,975 |   249,684
        FR |         0    22,139    37,086    51,104    65,957    66,440    69,242    72,982    74,410    77,889    54,199    40,321    22,616 |   654,385
        HR |         0         0         0         0         0         0         0     9,863    12,642    11,807    11,140    11,344     9,800 |    66,596
        HU |         0         0    14,567    20,291    22,582    22,685    24,732    24,887    29,729    28,888    26,064    23,225    13,226 |   250,876
        IE |         0     8,300    11,729    10,787    10,060    11,948    10,852     7,146     8,791    10,195    10,226     5,974         0 |   106,008
        IS |         0     6,128     8,259     7,381     7,138     7,103     7,142     7,590     7,766    14,715    12,617    10,875     6,092 |   102,806
        IT |         0    46,629    57,124    55,357    53,534    53,272    51,902    48,010    49,070    48,410    45,201    47,803    32,912 |   589,224
        LT |         0         0     9,100    12,448    13,072    12,379    13,136    13,468    12,879    13,005    12,023    12,118     8,781 |   132,409
        LU |     7,861     7,987    14,366    21,329    31,016    34,744    35,277    28,650    21,992    16,394    10,174    10,117     6,417 |   246,324
        LV |         0         0     9,018    11,260    11,468    13,451    14,807    15,856    16,363    15,723    15,107    14,424    10,180 |   147,657
        MT |         0         0         0     3,376     6,219     8,027    10,309    10,533    11,347    12,103    12,138    12,012     8,444 |    94,508
        NL |         0         0    21,634    23,371    26,202    25,739    23,973    24,916    25,758    25,255    25,019    24,845    16,497 |   263,209
        NO |    11,662    21,315    30,344    36,023    36,235    34,489    32,300    29,670    25,571    23,706    19,485    18,312    11,246 |   330,358
        PL |         0         0    36,525    46,022    43,629    42,071    39,365    38,056    37,448    37,861    37,153    36,786    26,027 |   420,943
        PT |         0     7,092     9,921    12,251    11,857    11,985    13,175    13,576    14,860    16,229    16,659    17,469    12,894 |   157,968
        RO |         0         0         0         0    14,902    19,219    18,844    18,422    18,047    17,777    17,707    17,409    12,838 |   155,165
        RS |         0         0         0         0         0         0         0         0         0         0    15,175    19,345    13,694 |    48,214
        SE |         0    13,734    19,143    16,240    16,358    16,205    16,072    15,688    14,150    14,665    14,279    14,313    10,006 |   180,853
        SI |         0         0    27,679    31,903    28,885    29,492    30,179    30,127    29,337    28,634    27,786    28,190    17,818 |   310,030
        SK |         0         0    11,779    15,140    14,188    15,225    15,865    15,760    15,454    15,539    15,569    15,734    11,895 |   162,148
        UK |         0         0    35,752    40,797    22,278    21,463    19,720    19,060    18,948    26,346    26,205    15,753         0 |   246,322
-----------+------------------------------------------------------------------------------------------------------------------------------------+----------
     Total |    23,615   222,865   479,509   567,098   597,547   614,833   618,169   620,467   619,632   630,191   602,729   575,213   380,736 | 6,552,604
```

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.

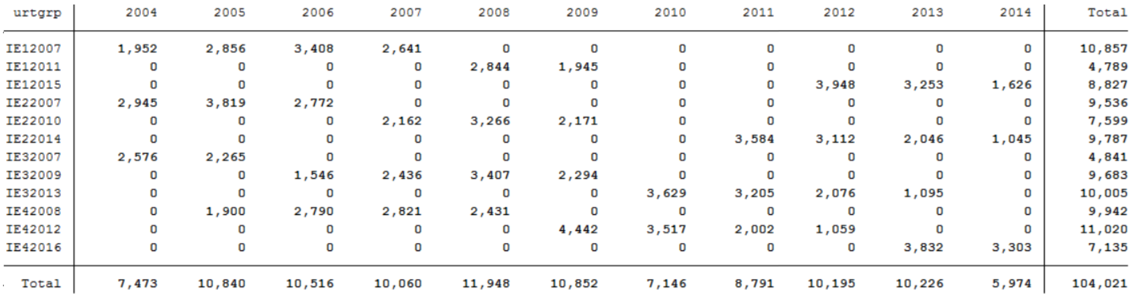


```
   urtgrp |     2004     2005     2006     2007     2008     2009     2010     2011     2012     2013     2014     2015 |     Total
----------+------------------------------------------------------------------------------------------------------------+----------
  AT12008 |        0    4,818    4,225    3,827    2,785        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 |    15,278
  AT12016 |        0        0        0        0        0        0        0        0        0    4,079    3,259    2,963 |    10,301
  AT22007 |    2,824    2,155        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 |    15,138
  AT22013 |        0        0        0        0        0        0    4,450    3,817    3,480    3,090        0        0 |    14,837
  AT22017 |        0        0        0        0        0        0        0        0        0        0    4,247    3,633 |     7,880
  AT32007 |    2,910    2,166    1,998        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 |    14,957
  AT32014 |        0        0        0        0        0        0        0    4,080    3,457    3,044    2,739        0 |    13,320
  AT42007 |    5,816    4,212    3,926    3,738        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 |    13,503
  AT42015 |        0        0        0        0        0        0        0        0    4,056    3,306    2,937    2,790 |    13,089
----------+------------------------------------------------------------------------------------------------------------+----------
    Total |   11,550   13,351   15,113   17,035   13,935   13,882   14,373   14,184   14,193   13,519   13,182    9,386 |   163,703
```

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 2014 to obtain a smaller, but potentially more representative sample.

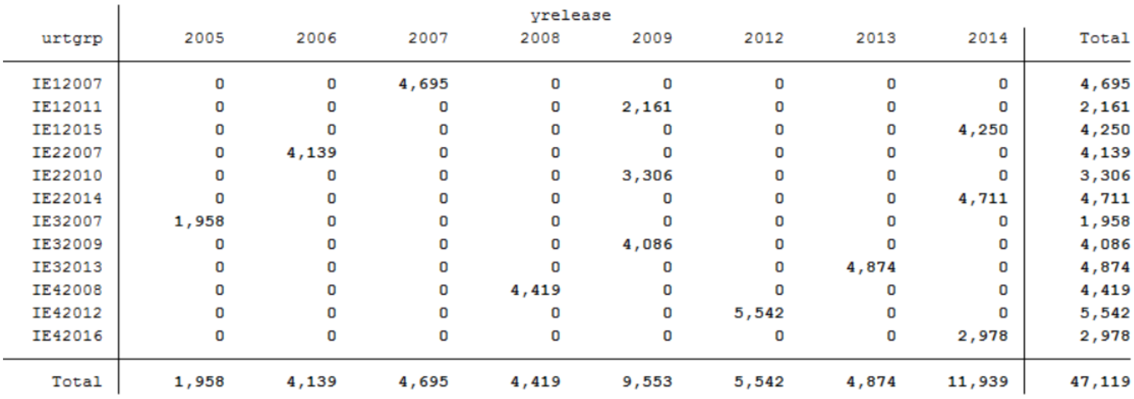
`tab urtgrp year if country == “IE”` shows the rotational groups of Ireland. In this case, one can see some irregularities: 2010 contains only two rotational groups, while 2011 contains three.



```
   urtgrp |     2004     2005     2006     2007     2008     2009     2010     2011     2012     2013     2014 |     Total
----------+---------------------------------------------------------------------------------------------------+----------
  IE12007 |    1,952    2,856    3,408    2,641        0        0        0        0        0        0        0 |    10,857
  IE12011 |        0        0        0        0    2,844    1,945        0        0        0        0        0 |     4,789
  IE12015 |        0        0        0        0        0        0        0        0    3,948    3,253    1,626 |     8,827
  IE22007 |    2,945    3,819    2,772        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 |     7,599
  IE22014 |        0        0        0        0        0        0        0    3,584    3,112    2,046    1,045 |     9,787
  IE32007 |    2,576    2,265        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 |     9,683
  IE32013 |        0        0        0        0        0        0    3,629    3,205    2,076    1,095        0 |    10,005
  IE42008 |        0    1,900    2,790    2,821    2,431        0        0        0        0        0        0 |     9,942
  IE42012 |        0        0        0        0        0    4,442    3,517    2,002    1,059        0        0 |    11,020
  IE42016 |        0        0        0        0        0        0        0        0        0    3,832    3,303 |     7,135
----------+---------------------------------------------------------------------------------------------------+----------
    Total |    7,473   10,840   10,516   10,060   11,948   10,852    7,146    8,791   10,195   10,226    5,974 |   104,021
```

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

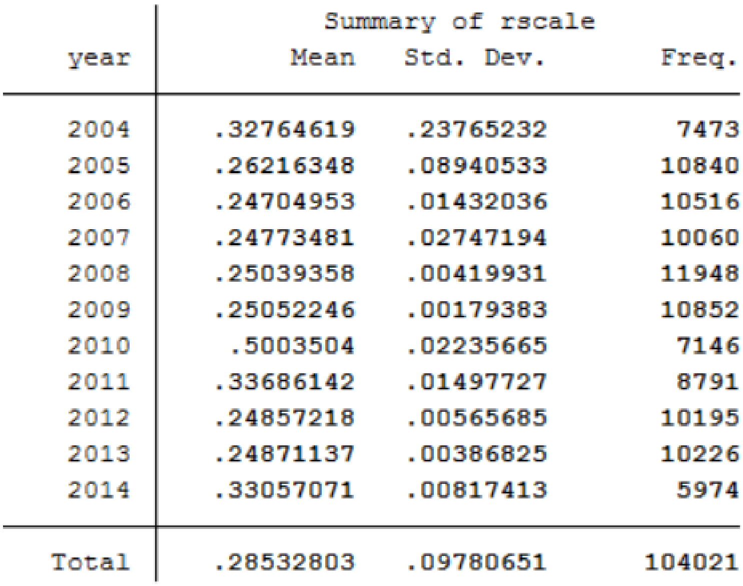
What happened? By using masterD.dta and running `tab urtgrp yrelease if country == “IE”`, it turns out that no data has been collected from Ireland in the 2010 and 2011 release because Ireland didn't contribute to these releases.



```
           |                                  yrelease
    urtgrp |      2005       2006       2007       2008       2009       2012       2013       2014 |     Total
-----------+----------------------------------------------------------------------------------------+----------
   IE12007 |         0          0      4,695          0          0          0          0          0 |     4,695
   IE12011 |         0          0          0          0      2,161          0          0          0 |     2,161
   IE12015 |         0          0          0          0          0          0          0      4,250 |     4,250
   IE22007 |         0      4,139          0          0          0          0          0          0 |     4,139
   IE22010 |         0          0          0          0      3,306          0          0          0 |     3,306
   IE22014 |         0          0          0          0          0          0          0      4,711 |     4,711
   IE32007 |     1,958          0          0          0          0          0          0          0 |     1,958
   IE32009 |         0          0          0          0      4,086          0          0          0 |     4,086
   IE32013 |         0          0          0          0          0          0      4,874          0 |     4,874
   IE42008 |         0          0          0      4,419          0          0          0          0 |     4,419
   IE42012 |         0          0          0          0          0      5,542          0          0 |     5,542
   IE42016 |         0          0          0          0          0          0          0      2,978 |     2,978
-----------+----------------------------------------------------------------------------------------+----------
     Total |     1,958      4,139      4,695      4,419      9,553      5,542      4,874     11,939 |    47,119
```

Figure 7: tab urtgrp yrelease if country == "IE", Stata page print, masterD.dta

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.



```
            |       Summary of rscale
       year |        Mean   Std. Dev.       Freq.
------------+------------------------------------
       2004 |   .32764619   .23765232        7473
       2005 |   .26216348   .08940533       10840
       2006 |   .24704953   .01432036       10516
       2007 |   .24773481   .02747194       10060
       2008 |   .25039358   .00419931       11948
       2009 |   .25052246   .00179383       10852
       2010 |    .5003504   .02235665        7146
       2011 |   .33686142   .01497727        8791
       2012 |   .24857218   .00565685       10195
       2013 |   .24871137   .00386825       10226
       2014 |   .33057071   .00817413        5974
------------+------------------------------------
      Total |   .28532803   .09780651      104021
```

Figure 8: tab year , summarize (rscale) , Stata page print, masterR.dta (Ireland only)

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.

5 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

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

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? 2014) 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 8 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. 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, France is the only country in which rotation groups are cut and pieced together across different releases. This makes it impossible to use the longitudinal weight *rb064s* with France. 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. *HB100* (minutes of time needed to complete the questionnaire is a string with some non-numeric characters in 2011 release. De-string and set nonnumeric characters as missing.