1 Data – EU-silc

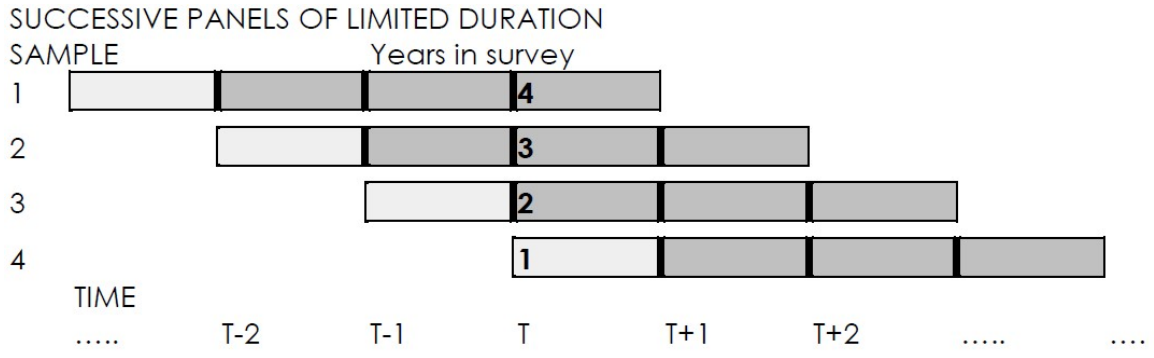
This is a short guide on how to build a panel dataset for the years 2003-2013 based on the UDB files issued by Eurostat. It is accompanied by a Stata do-file (buildmasterfiles.do) and its results when applied to the EU-silc UDB files (masterD.dta, masterH.dta, master.dta, masterP.dta)

1.1 Essential practical information about the EU-silc dataset(s)

The longitudinal data for EU-silc is collected following an "integrated" or "rotational" design (p. 16 Guidelines). 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, as the scheme below shows. The reason behind this choice is that the integrated design minimizes practical issues 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 (referred to as "rotational groups"), that are still "active". Looking at the scheme below, 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. Once Eurostat has received the data from all countries, it merges them and distributes them without adding any data from previous years.

This means that in order to have a panel with roughly the same number of observations each year, it is necessary to add observations contained in previous rotational groups, that are not "active" anymore, and therefore located in files distributed during previous years. Obviously, we can also build datasets that cover a time span that goes beyond four years, even though this is not straight forward since the rotational groups are not identified homogenously across countries with respect to their stage. For example, rotational group "4" in the scheme might be identified as rotational group "3" in Italy, but in Greece it is labeled as "2".



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

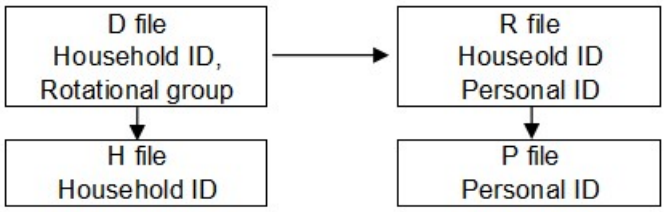
1. UDB_l13**D**_ver 2013-2 from 01-01-2016.csv

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,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 house-hold level during the surveys, such as total gross household income, housing costs and so on. Obviously there are some households that are present in the register, but then didn't participate in the surveys, so they are missing in the D file.
3. The *R* file is the register file relating to the individual surveys. 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, such as occupation, salary and so on.

Within one release household and individual datasets are linked between each other through IDs that indicate the households 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*) can be found again in the *H* file. In the *R* we may find the individuals being part of that household, i.e. the individuals that share the same *hid*, and in the *P* file are individuals from the *R* file.



Note how the information regarding the rotational group is contained only in the *D* file. This means, that if we want select certain individual data contained in the P file pertaining to one rotational group, we must merge the *D*, *R* and *P* file to gather all the necessary information in terms of IDs.

In terms of numbers of observation, not all households listed in the register file (*D*) file, are also listed in the *H* file and their members might be missing from the *R* file. Furthermore, not all individuals contained in the *R* file are also in *P*.

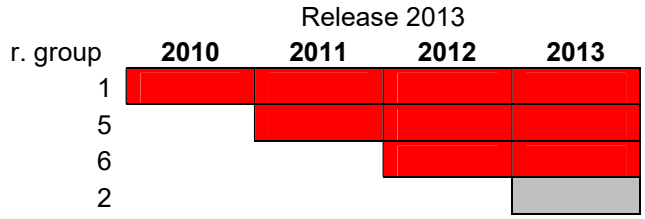
The household and individual IDs are unique within each country and within one release. On the other hand, they are not unique across countries, and across rotational groups that are not part of the same release. Also rotational groups don't have an ID that uniquely defines them across different releases within the same country – thus, for example, a country may call a rotational group containing 4 years of observations "3" in the 2010 release and in the 2012 release the new sub-sample, which takes the place of the old one is identified by "3" again, even if they contain different households. Furthermore, within a sub-sample sometimes data collected and distributed in previous years is revised. These

are reasons enough to understand why it is not possible to simply copy the data for a given year from an older release and then merge with the last one.
Finally, a word of caution: the complexity as well as the number of different sources makes EU-silc prone to inconsistencies (check the "Problems and Modifications" file attached to each release for details and also see the "Open first" document). The approach that follows works under the assumption that all countries actually followed the guidelines provided by Eurostat and handed in a dataset that respects all requirements listed under "Transmission of Data and Data Availability" in the Guidelines (p.54), which in reality is not always the case. Thus, running the script without verifying manually each country-specific dataset and not being aware of problems in each release may result loss of observations and errors. It is also necessary to check for inconsistencies and changes in the variables across releases before starting a statistical analysis of the data. Croatia (country code *HR*), for instance, didn't respect the requirements for the 2011 and the 2012 release, so it would lose roughly 10% of observations during the process described below without tweaking the script (even though one should try to understand first why a given country didn't respect the requirements). At the end of the document is a non-exhaustive list of issues we came across during the process of writing the script, and the solutions that have been adopted.

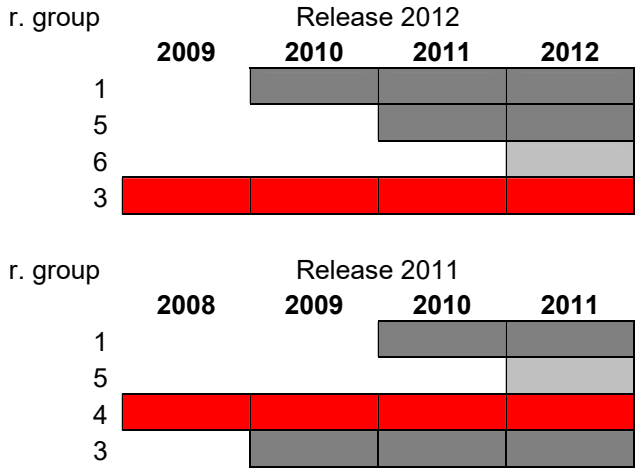
1.2 Building the EU-silc longitudinal dataset 2003-2013

The aim is to write a Stata script (buildmasterfiles.do) that starts from the latest release (in this case 2013) and then adds to the file gradually the data from previous years by merging the last release with data from those sub-samples that are not active anymore.

To understand the mechanism in which the data from different releases should be merged, consider the following: if a country has been taking part in EU-silc from 2008, its 2013 release should have the following the structure in terms of rotation groups:



Notice how group number 2 is not contained in the dataset (light grey box). This will eventually lead to the fact that after having merged data from all years, there will be a drop in observations from 2012 to 2013.
To add further data to the 2013 release, we want to go to the 2012 release and select the rotation group which was completed in 2012 (group 3), and which would be replaced by the new group 2 in 2013. The selection is done by checking which rotation group contains households with most observations over the years. Next, one would open Release 2011 and get group 4 and so on moving back in time.



To select the groups of interest, one could simply open each release and check which group has households with four observations. The Problem with this approach, however, is that not all countries entered EU-silc at the same time. Thus, 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. Thus, continuing the example from above, if a country started participating in EU-silc in 2008, the selection process would be as follows: the maximum number of observations of households would be 3 in the 2010 release, but only group 7 is selected because group 4 has already been added to the dataset with release 2011. In the 2009 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 upon comparing the unique identifier of the rotational group *urtgrp* between different releases, which is preferable to the unique household *ID* (*uhid*) because sometimes observations that had already been collected in earlier releases are dropped in the current release, even though they should be there. 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 want to not select them because we assume there is a good reason they haven't been reported in the more recent release.

1.3 How the script works

This paragraph gives a short illustration on how the script works; first the *D* file from the 2013 release is loaded. The variable names in EU-silc are codes, soothe first step is always to rename the most important variables to *year*, *country*, *rotation_group*, *hid* and *pid*, where the latter two are household and personal IDs respectively.

Next, the script identifies, which rotational groups in each country contain four and less observations and therefore are going to be dropped during the 2014 release. After that, going backwards in time, it proceeds to past releases to select the rotational groups with the most observations in time. The results are saved in 20XXD.dta and 20XXslctd.dta files, where the former are used to build the *D* file (masterD.dta) for the overall panel while latter contain only essential information and are used to check whether rotation groups form more distant releases have already been selected. This is done by merging the 20XXslctd.dta files with the release and dropping the matches. During that process the script also gives each household a unique ID, *uhid*, based on the country, rotational group, year in which the rotational group is dropped and household ID. At the end of all 20XXD.dta files are merged into one masterD.dta file.

Going on, the script proceeds to the *H* files by selecting those observations that are contained in the rotation groups identified in the 20XXslctd.dta files by merging and matching. As before, the results from each release are saved and then merged to one file, masterH.dta.

After that the script opens the personal registers, i.e. the *R* files. The procedure is essentially the same as the one with the *H* files, only that also a unique personal ID *upid* is generated the selected individuals from each release are saved in the 20XXslctdR.dta files. The result of this procedure is the masterR.dta file.

Finally, the script moves on to the personal data, the *P* files. Here the subgroups are selected based on the personal ID and not the household ID, since it is not contained in these files. The script ends generating masterP.dta and erasing all superfluous files that have been generated during the process.

1.4 Issues and tweaks

As already pointed out, EU-silc presents some issues. Here is a list of inconsistencies that became clear from running the script, and the respective work-arounds. There may be more obviously.

1. Greece: the variable country at the beginning is GR, then changes to EL in more recent releases starting from 2008. Changed it to EL in all releases, but haven't checked whether there are inconsistencies with the regions as well.
2. Croatia: 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 selects all data from the 2013 release and

discards the rest. In particular the observations printed in bold are lost. In the absence of information on the cause of these issues, it is probably better not to intervene, since there is no solution without making some strong assumptions.

HR, 2013 release

Group	1	2	4
Year			
2010	0	3,504	0
2011	0	1,844	3,496
2012	3,483	1,549	1,754
2013	1,765	1,375	1,506

HR, 2012 release

group	1	2	3	4
year				
2010	**1,869**	1,834	0	0
2011	**1,887**	1,845	**1,739**	1,746
2012	0	1,550	**1,743**	1,754

HR, 2011 release

group	1	2
year		
2010	**3,489**	3,505
2011	**1,887**	1,845

3. Romania (RO) lacks from the 2013 release, Germany from all releases.
4. Some countries present duplicates in the *H* file, or at least *hid* is not a unique identifier. The number of observations is relatively small (200 – 400 each release compared to ca. 50.000 observations), so they can be dropped.
5. In some countries (*NL, EE*) there are households contained in *H* file, but absent from the register (*D*). Since we don't know which rotational groups they belong to, they must be dropped.
6. *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.
7. Stata 12 seems to have problems when it comes to working with IDs made of many numbers. Transform IDs from numbers into strings.
8. The *R* files contain several thousand duplicates in each release. Compared to more than 1 m observations they are not too many. They are probably duplicates in a proper sense (i.e. all data is identical, not only IDs).