CSB
WORKING PAPER

centreforsocialpolicy.eu June 2016 No 16 / 02

Notes on updating the EU-SILC UDB sample design variables 2012-2014

Lorena Zardo Trindade & Tim Goedemé

Universiteit Antwerpen

CSB

University of Antwerp
Herman Deleeck Centre for Social Policy
Sint-Jacobstraat 2
B – 2000 Antwerp
fax +32 (0)3 265 57 98

Notes on updating the EU-SILC UDB sample design variables 2012-2014

Lorena Zardo Trindade & Tim Goedemé

Working Paper No. 16 / 02

June 2016

ABSTRACT

Indicators based on EU-SILC should be accompanied by appropriate standard errors, and in order to do so, it is necessary to consider the sample design. Because important sample design variables are missing in the EU-SILC User Database (UDB), the aim of this note is to explain the update of EU-SILC UDB sample design variables for 2012 (version 4), 2013 (version 3) and 2014 (version 1), based on the methodology developed by Goedemé (2010b, 2013a). Although several of the challenges for reconstructing the EU-SILC sample design variables are identical for all releases of the data, the update required minor adjustments in the computation of the new sample design variables psu1 and strata1. The effect of the use of the new sample design variables on standard errors is observed for 'At risk of poverty' and 'Material deprivation', for which SE values are found to be larger when the sample design variables are considered.

Keywords: EU-SILC, sample design, standard error, sampling variance

Corresponding author:
Tim Goedemé
Herman Deleeck Centre for Social Policy (CSB)
Faculty of Political and Social Sciences
University of Antwerp
Tel: +32 3 265 55 55
Email: tim.goedeme@uantwerpen.be

1 Introduction

This note aims to explain the update of EU-SILC UDB sample design variables for 2012 and 2013, based on the methodology developed by Goedemé (2010b, 2013a). We briefly recall the main points of the discussion.

Indicators based on EU-SILC should be accompanied by appropriate estimates of their precision and statistical reliability, and in order to do so, it is necessary to consider the sample design design (e.g. Kish 1965; Heeringa et al., 2010). Unfortunately, important sample design variables are missing in the EU-SILC User Database (UDB). Previous studies have shown that neglecting the sample design can lead to an underestimation of standard errors. Focusing on poverty indicators, among others, Rodgers and Rodgers (1993), Howes and Lanjouw (1998) and Jolliffe et al. (2004) measured the impact of sample design variables on estimated standard errors. More recently, Goedemé (2010b, 2013a, 2013b) discusses EU-SILC sample design variables and their importance to the estimation of Europe 2020 poverty indicators. In addition, he proposed a procedure for reconstructing the sample design variables, given that these variables are lacking in the EU-SILC UDB for most EU-SILC countries.

EU-SILC UDB sample design variables based on Goedemé (2010b, 2013a) must be updated at every new release of the survey. Information available on EU-SILC datasets is frequently modified due to changes in sample design, sample frame, availability and quality of sample design variables in the UDB and inclusion of new countries. Currently, much of this information is reported in the national quality reports as well as the comparative quality report. However, in some cases important information is lacking if one wants to compute sampling variances using the available data.

In order to describe the update of EU-SILC UDB sample design variables for 2012, 2013 and 2014 this note is structured as follows. Section two briefly describes the EU-SILC sample design. Section three discusses the issues confronted while reconstructing the EU-SILC sample design variables and the changes in the syntax for 2012 (version 4), 2013 (version 3) and 2014 (version 1) when compared to 2011. In addition, the constructed sample design variables are analyzed and the effect of the use of sample design variables on standard errors is illustrated. Finally, the fourth section summarizes the main points of the note and an example of the Stata do-files used to compute the construction of strata and PSU variables is included in the annex.

2 EU-SILC sample design

Sample design, sample frame and data source are substantially different among EU-SILC member states. Although EU-SILC uses harmonized methods and definitions in order to establish reliable comparisons between EU Member States, there are considerable differences in sample design between EU-SILC countries (Table 1). In some countries, single stage designs are in use, whereas in other countries two- or three stage designs are employed. Most countries apply stratification for at least one stage. Both sampling with equal probabilities of selection and sampling with probabilities proportional to size are in use and in some cases systematic sampling is applied. Sample frames range from censuses to different kinds of population registers, and procedures for updating the sample frame as well as the date of the last update are not standardized. Sample frames from Germany and the Netherlands are special cases that can be source of concern. For the German data, the sample frame consists of households included in the Microcensus that have indicated that they are willing to participate in additional surveys; for the Netherlands, the sample frame consists of households who have successfully participated in several waves of the Labour Force Survey (LFS) (Goedemé, 2010a and 2010b).

EU-SILC has a 4-year rotational panel design for the great majority of EU countries. In several countries the rotational period is longer. In a four-year rotational design, sample units remain in the sample for four years (waves) and one quarter of the sample is replaced each year. Each quarter of the sample is known as a replication, and each replication is representative of the target population. In France, for instance, the panel rotation is spread across a longer time period (9-year panel), and in Luxemburg a

Table 1: EU-SILC sample design by country, 2010-2013

Sample design		Country
Without stratification	Simple random sampling	MT, DK, IS, NO
	Systematic sampling	SE
With stratification	Stratified simple random sampling	LU, CY, SK, CH, LT, DE, AT
	Stratified and systematic sampling	EE
	Stratified two-stage	PT, SI, NL, HR, IT, LV
	Stratified multi-stage	CZ, ES, PL, RO, IE, FR, UK, BE, BG, EL
	Stratified two-phase	FI, HU

Source: Berger, Osier and Goedemé (forthcoming), based on Eurostat, 2012 EU-SILC Comparative Quality Report (available on CIRCABC).

pure panel is implemented (Goedemé, 2013b). In most countries rotation is implemented at the level of primary sampling units (PSUs, clusters of households selected at the first stage of the sample selection process). However, in some others (e.g. BE and ES), rotation is implemented at the level of households and PSUs remain the same for the entire duration of EU-SILC. When estimating the statistical reliability of a change over time, it is important to take the covariance between the waves that are compared into account: given that part of the sample remains the same, estimates about one wave tend to be correlated with those of another wave. In practice, this requires that the identification of households and PSUs is consistent across the waves compared. Unfortunately, this is currently not the case in the EU-SILC UDB. In other words, when comparing across waves, it might be that some covariance is estimated, based on coincidental similarities in PSU and household identifiers in the UDB. Therefore, we suggest to make sure sample design variables are unique across waves when estimating the sampling variance (or standard errors, p-values and confidence intervals) of changes over time. As a result, one can be sure the sampling variance of the change over time will be over-estimated, without being confounded by incidental similarities in PSU codes and household identifiers.

Accounting for the sample design while estimating standard errors usually requires two different types of information: (i) a complete description of the implemented sample design; and (ii) accurate variables that describe stratification and clustering. If strata are not taken into account, confidence intervals can be overestimated and the researcher might be unduly conservative. In addition, if clustering is neglected, standard errors will be underestimated, and relations which are not statistically significant may appear to be significant. However, in the case of the EU-SILC UDB, lack of detailed information on sample design strategies and inaccurate sample design variables (strata and clusters) impose certain constraints for properly estimating the standard errors. When calculating standard errors in regular software packages such as Stata, only the first stage is taken into account (the so-called 'ultimate cluster approach', see Osier (2012), Heeringa et al. (2010), Wolter (2007), and Kalton (1979)). This implies that accurate information about the first stage of the sample is required for the estimation of standard errors. For this purpose, variables identifying the primary sampling units and the strata used at the first stage should be suitable. Given that the stratification variable (DB050) is lacking in the EU-SILC UDB and the national quality reports only provide limited information about stratification criteria and the number of strata (Table 2), some manipulation of variables is required to reconstruct suitable sample design variables.

Table 2: Criteria for first level of stratification and reported number of strata by country

Country	Stratification criteria	Reported number of strata[1]
AT	Geographic region	207
BE	Geographic region	11
BG	Geographic region	56
CH	NA	NA
CY	Geographic region	9
CZ	Geographic region and population size	53[2]
DE	Geographic region and socio-economic standards	NA
DK	-	1
EE	Geographic region	3
EL	Geographic region and population size	90
ES	Geographic region and population size	93[2]
FI	Socio-economic categories	13
FR	Geographic region and degree of urbanization	88[2]
HR	NA	NA
HU	Geographic region and degree of urbanization	NA
IE	NA	NA
IS	-	1
IT	Geographic region	NA
LT	Degree of urbanization	7
LU	Socioeconomic categories	9
LV	Degree of urbanization of the area	4
MT	-	1
NL	Geographic region	40
NO	-	1
PL	Geographic region	250
PT	Geographic region	7
RO	Geographic region and degree of urbanization	88
RS	NA	NA
SE	-	1
SI	NA	NA
SK	Geographic region and degree of urbanization	48
UK	Geographic region	31

Note: (1) The reported number of strata is based on information from the newest National Quality Report available, unless indicated otherwise. (2) Assuming no changes in sample design, information is based on National Quality Reports prior to 2012. (-) indicates the cases in which stratification was not applied. "NA" refers to Not Available.
Source: National Quality Reports.

3 Construction of the sample design variables

Several issues had to be confronted for reconstructing the EU-SILC sample design variables, and many of them are identical for all releases of the data. For this reason, most of the issues identified for the 2012, 2013 and 2014 waves have already been extensively discussed elsewhere, along with detailed guidelines and recommendations for the construction of the variables for primary strata and primary sampling units (PSUs) (see in particular Goedemé, 2010b and 2013b).

1. **EU-SILC UDB does not include the original stratification variable (DB050).** For confidentiality reasons, the original stratification variable (DB050) is not included in the EU-SILC UDB. For a number of countries, as proxy to the stratification variable, one could use DB040 (Belgium, Czech Republic, Greece, Spain, France, Italy, Romania). This variable contains the region of residence at the moment of the interview, instead of the stratum at the moment of selection. Households that have moved between the moment of selection and the moment of interview should be assigned to their original stratum of selection. For quite a few countries, the number of strata is seriously

under-estimated. Extreme cases refer to countries with stratified samples in which DB040 is missing (Germany, Netherlands, Portugal and Slovenia) (Table 3). It should be noted that if this variable (and its flag variable) would be included in the EU-SILC UDB, many of the problems identified below could also be avoided.

2. **PSU variable (DB060) is missing or not properly coded for some countries.** In several countries DB060 is missing or not properly coded even though the sample has been clustered on a higher level than the household level. For Hungary, DB060 is partially lacking in 2012, 2013 and 2014. In the case of France, DB060 is partially lacking in 2012 and 2014 and completely missing in 2013 (Table 3).[1] For both situations, we advise that PSUs must be identified by household ID (DB030) or DB062, when available.

3. **PSUs (DB060) are often not unique across strata.** In countries like Bulgaria and, Poland, DB060 is not unique across strata. With an inadequate stratification variable, this implies that PSUs that belong to different strata would be taken together if DB040 would be used as a stratification variable and DB060 as PSU variable. Therefore, we recommend in these cases to ignore stratification, and only take account of clustering at the household level, rather than using DB060 as PSU variable. Especially in this case, further analysis of the dataset with complete sample design information would be very useful.

4. **Some PSUs are self-representing. In some countries, certain PSUs are always included in the EU-SILC samples (e.g. the biggest cities), regardless of its rotational aspect.** Such PSUs are often described as 'self-representing', and they are included in the sample with a probability of selection equal to 1. Therefore, self-representing PSUs should be treated as a separate stratum for variance estimation purposes. Italy, the United Kingdom (one self-representing PSU) and France are examples of countries that include self-representing PSUs in their samples. If self-representing PSUs are treated as regular PSUs instead of strata, it might in some cases result in a considerable over-estimation of the sampling variance. Ideally, in the case of self-representing PSUs, information about stratification at the second stage of the sample design is required.[2] However, currently, such a variable does not exist in the dataset. By lack of this information self-representing PSUs should be considered stratum rather than PSU in cases they can be identified in the data files. For Italy, self-representing PSUs can be identified as those PSUs with households from several panels, rather than one. In the case of France, until 2009, 53 PSUs with information on secondary sampling units (DB062) and with the largest weighted number of households were assumed to be self-representing (i.e. they were considered strata and DB062 is used as PSU variable). However, changes in sample design in 2010 complicated the matter and the method was no longer satisfactory for identifying self-representing PSUs. In the case of the United Kingdom, the self-representing PSU could be identified in a similar way as was the case for Italy.

5. **PSUs (DB060) can be split across strata if DB040 is used as a proxy for stratification. Given the panel character of EU-SILC, households may move from one region to another between the moment of selection and the moment of interview.** This results in PSUs being 'split' across various strata. This can be the case for countries for which DB040 has been used as stratum identifier (Belgium, Czech Republic, Greece, Spain, France, Italy, Romania). In these circumstances, split PSUs should be re-allocated with a reasonable degree of certainty to the correct stratum (in practice, the region inhabited by the majority of households of the PSU to which they belong). As a result, in the UDB households belonging to different strata but with the same PSU code are treated as one PSU and allocated to one region identified by DB040.

[1] Because the Quality Report 2013 is not yet available for France, the accuracy of DB060 cannot be confirmed.

[2] Goedemé (2010b) suggests that self-representing PSUs and their substrata at the second stage of the sample selection scheme could be immediately coded as primary strata, and the sampling units at the subsequent stage of the sample design as primary sampling units. Please note that since SILC 2013 for a selection of countries and SILC 2014 for all countries, the coding of the PSUS variables and their flags has improved. Guidelines follow now the recommendations made in Goedemé (2013b).

6. **Systematic sampling is ignored.** For reasons explained elsewhere (e.g. Goedemé, 2013b), we do not take the order of selection of PSUs into account (DB070) for defining computational strata and PSUs. This may lead to an overestimation of the sampling variance in countries with systematic selection on an ordered sampling frame.

Table 4 describes the sample design variables constructed for EU-SILC UDB from 2010 to 2014 using the methodology developed by Goedemé (2010b, 2013a). According to the description, six countries had significant changes in the number of PSUs and/or number of observations. In the case of France, the number of PSUs has increased from 1,205 in 2012 to 11,090 in 2013, and decreased to 1,049 in 2014. which is explained by the fact that household identifiers were considered as PSUs since DB060 is not available for EU-SILC UDB 2013, containing only six unique identifiers). For Portugal, the number of PSUs has increased from 542 in 2012, to 1,994 in 2013 and to 2,277 in 2014, which ca be explained by changes in the sampling frame between 2012 and 2013. According to the countries quality report, from 2004 to 2012 the primary sampling units (PSU) were the areas of a master sample based on census enumeration areas. From 2013 onwards, the master sample is based on the National Dwellings Register, in which PSU are one constituted by one or more contiguous grid cells with 1 Km2 of area. Poland and Bulgaria had a significant increase in the number of PSUs from 2010 to 2011. Until 2010, both countries had been stratified by DB040. However, since DB060 was not unique across strata, the number of PSUs was underestimated. Because of that, from 2011 onwards, the samples for Poland and Bulgaria were no longer stratified and household identifiers have been considered as PSUs. For Luxembourg, the number of observations (households) has decreased between 2012 and 2013, from 6,031 to 3,770. According to the country's quality report, this change reflects the reduction of the achieved sample size (i.e. the number of observed sampling units with an accepted interview), as the actual sample size (i.e. the number of sampling units selected in the sample) was 7,427. Significant changes in the number of observations were also observed for Croatia. The country's sample experienced an increase in the number of total observations and PSUs between 2010 and 2011. However, unlike Luxembourg, this increase can be explained by an increase of the actual sample size. The Netherlands had a significant increase in the number of PSUs from 435 in 2013 to 10,174 in 2014. Although there have been no changes in sample design, DB060 has unique values for all observations in 2014, which makes the case of the Netherlands similar to those countries for which household identifiers are considered as PSUs.

The sample design for the United Kingdom has changed significantly for the 2012, 2013 and 2014 releases. Before 2012, all households for the cross-sectional and longitudinal EU-SILC originated from the General Lifestyle Survey (GLF) sample. Since then, cross-sectional and wave 1 respondents from the longitudinal panel have been selected from the Family Resources Survey (FRS). The total number of observations (households) has increased for Great Britain, while it decreased for Northern Ireland. In addition, the number of PSUs for Northern Ireland has decreased when comparing 2012, 2013 and 2014 to the previous years.

Several changes have been implemented with regard to stratification. In Greece DB040 is available since EU-SILC 2011, and can now be used as a proxy. In the case of Bulgaria and Poland, we do no longer use DB040 as a proxy for stratification, given that DB060 is not unique across strata. We consider it a better proxy to simply use household IDs as PSUs. However, for the latter two countries it is impossible to say whether this results in an underestimation or over-estimation of the sampling variance. Some indications are available in Goedemé (2013a), though. A new comparison with estimates on the basis of the complete sample design variables could shed more light on this./footnote(For EU-SILC UDB 2012 (version 3) and 2013 (version 2), DB040 was missing for Belgium and could not be used as a proxy for regional stratification. When using these versions, the Belgium sample could be treated as a simple random sample of PSUs. However, it is important to note that this most probably leads to an overestimation of the variance.)

Table 3: EU-SILC UDB sample design information (D-File), 2012-2014

Country	Year	Missing observations						Unique identifiers						Total n. obs.
		DB040	DB060	DB062	DB070	DB075	DB030	DB040	DB060	DB062	DB070	DB075	DB030	
AT	2012	0	6,232	6,232	6,232	0	0	3	0	0	0	4	6,232	6,232
	2013	0	5,977	5,977	5,977	0	0	3	0	0	0	4	5,977	5,977
	2014	0	5,909	5,909	5,909	0	0	3	0	0	0	4	5,909	5,909
BE	2012	0	0	5,817	0	0	0	3	274	0	244	4	5,817	5,817
	2013	0	0	6,159	0	0	0	3	275	0	235	4	6,159	6,159
	2014	0	0	6,021	0	0	0	3	275	0	19	4	6,021	6,021
BG	2012	0	0	0	5,706	0	0	2	654	5	0	4	5,706	5,706
	2013	0	0	0	4,971	0	0	2	622	5	0	4	4,971	4,971
	2014	0	0	0	4,963	0	0	2	635	5	0	4	4,963	4,963
CH	2012	0	7,529	7,529	7,529	0	0	1	0	0	0	4	7,529	7,529
	2013	0	7,341	7,341	7,341	0	0	1	0	0	0	4	7,341	7,341
	2014													
CY	2012	0	4,638	4,638	4,638	0	0	1	0	0	0	4	4,638	4,638
	2013	0	4,648	4,648	4,648	0	0	1	0	0	0	4	4,648	4,648
	2014	0	4,294	4,294	4,294	0	0	1	0	0	0	4	4,294	4,294
CZ	2012	0	0	8,773	8,773	0	0	8	1,661	0	0	4	8,773	8,773
	2013	0	0	8,275	8,275	0	0	8	1,589	0	0	4	8,275	8,275
	2014	0	0	8,053	8,053	0	0	8	1,551	0	0	4	8,053	8,053
DE	2012	13,145	13,145	13,145	13,145	0	0	0	0	0	0	4	13,145	13,145
	2013	12,703	12,703	12,703	12,703	0	0	0	0	0	0	4	12,703	12,703
	2014	12,744	12,744	12,744	12,744	0	0	0	0	0	0	4	12,744	12,744
DK	2012	0	5,355	5,355	5,355	0	0	1	0	0	0	4	5,355	5,355
	2013	0	5,419	5,419	5,419	0	0	1	0	0	0	4	5,419	5,419
	2014	0	5,758	5,758	5,758	0	0	1	0	0	0	4	5,758	5,758
EE	2012	0	5,433	5,433	0	0	0	1	0	0	5,303	4	5,433	5,433
	2013	0	5,775	5,775	0	0	0	1	0	0	5,504	4	5,775	5,775
	2014	0	5,871	5,871	0	0	0	1	0	0	1,544	4	5,871	5,871
EL	2012	0	0	0	0	0	0	4	1,094	28	32	4	5,626	5,626
	2013	0	0	0	0	0	0	4	1,340	23	38	4	7,439	7,439
	2014	0	0	0	0	0	0	4	1,536	23	48	4	8,620	8,620
ES	2012	0	0	12,714	12,714	0	0	19	1,996	0	0	4	12,714	12,714
	2013	0	0	12,139	0	0	0	19	1,989	0	104	4	12,139	12,139
	2014	0	0	11,965	0	0	0	19	1,989	0	104	4	11,965	11,965

(continuing)

(Table 3 continuation)

Country	Year	Missing observations						Unique identifiers						Total n. obs.
		DB040	DB060	DB062	DB070	DB075	DB030	DB040	DB060	DB062	DB070	DB075	DB030	
FI	2012	0	0	10,307	10,307	0	0	4	10,307	0	0	4	10,307	10,307
	2013	0	0	11,370	11,370	0	0	4	11,370	0	0	4	11,370	11,370
	2014	0	0	11,030	11,030	0	0	4	11,030	0	0	4	11,030	11,030
FR	2012	2	0	3,023	11,999	0	0	22	1,205	687	0	9	11,999	11,999
	2013	0	11,084	10,474	11,131	0	0	22	6	2	0	9	11,131	11,131
	2014	1	0	1,645	11,384	0	0	22	1,049	624	0	9	11,384	11,384
HR	2012	0	0	5,853	0	0	0	1	1,373	0	1,373	4	5,853	5,853
	2013	0	0	5,362	0	0	0	1	1,470	0	1,278	4	5,362	5,362
	2014	0	0	5,443	0	0	0	1	1,567	0	1,523	4	5,443	5,443
HU	2012	0	0	5,972	11,311	0	0	3	6,169	120	0	4	11,311	11,311
	2013	0	0	5,406	10,223	0	0	3	5,598	97	0	4	10,223	10,223
	2014	0	0	4,806	9,211	0	0	3	4,996	4,378	0	4	9,211	9,211
IE	2012	0	0	4,592	4,592	0	0	1	1,243	0	0	4	4,592	4,592
	2013	0	0	4,922	4,922	0	0	1	1,150	0	0	4	4,922	4,922
	2014	0	0	5,486	5,486	0	0	1	1,526	0	0	4	5,486	5,486
IS	2012	0	3,091	3,091	3,091	0	0	1	0	0	0	4	3,091	3,091
	2013	0	3,020	3,020	3,020	0	0	1	0	0	0	4	3,020	3,020
	2014	0	3,001	3,001	3,001	0	0	1	0	0	0	4	3,001	3,001
IT	2012	0	0	0	19,579	0	0	5	737	2,668	0	4	19,579	19,579
	2013	0	0	0	18,487	0	0	5	731	2,473	0	4	18,487	18,487
	2014	0	0	0	19,663	0	0	5	750	2,784	0	4	19,663	19,663
LT	2012	0	5,394	5,394	5,394	0	0	1	0	0	0	4	5,394	5,394
	2013	0	5,142	5,142	5,142	0	0	1	0	0	0	4	5,142	5,142
	2014	0	5,194	5,194	5,194	0	0	1	0	0	0	4	5,194	5,194
LU	2012	0	6,031	6,031	6,031	0	0	1	0	0	0	5	6,031	6,031
	2013	0	3,770	3,770	3,770	0	0	1	0	0	0	4	3,770	3,770
	2014	0	3,879	3,879	3,879	0	0	1	0	0	0	4	3,879	3,879
LV	2012	0	0	0	0	0	0	1	1,177	6,344	331	4	6,499	6,499
	2013	0	0	0	0	0	0	1	1,104	6,176	502	4	6,309	6,309
	2014	0	0	0	0	0	0	1	1,120	5,997	499	4	6,125	6,125
MT	2012	0	4,350	4,350	4,350	0	0	1	0	0	0	4	4,350	4,350
	2013	0	4,381	4,381	4,381	0	0	1	0	0	0	4	4,381	4,381
	2014	0	4,381	4,381	4,381	0	0	1	0	0	0	4	4,381	4,381

(continuing)

(Table 3 continuation)

Country	Year	Missing observations						Unique identifiers						Total n. obs.
		DB040	DB060	DB062	DB070	DB075	DB030	DB040	DB060	DB062	DB070	DB075	DB030	
NL	2012	10,168	0	0	10,168	0	0	0	439	4,404	0	4	10,168	10,168
	2013	10,131	0	0	10,131	0	0	0	435	5,369	0	4	10,131	10,131
	2014	10,174	0	10,174	10,174	0	0	0	10,174	0	0	4	10,174	10,174
NO	2012	0	6,050	6,050	6,050	0	0	1	0	0	0	4	6,050	6,050
	2013	0	6,031	6,031	6,031	0	0	1	0	0	0	4	6,031	6,031
	2014	2	7,371	7,371	7,371	0	0	1	0	0	0	4	7,371	7,371
PL	2012	0	0	0	13,116	0	0	6	112	254	0	4	13,116	13,116
	2013	0	0	0	12,899	0	0	6	119	283	0	4	12,899	12,899
	2014	0	0	0	12,978	0	0	6	5,436	282	0	4	12,978	12,978
PT	2012	6,257	0	6,257	0	0	0	0	542	0	133	4	6,257	6,257
	2013	6,491	0	6,491	0	0	0	0	1,994	0	561	4	6,491	6,491
	2014	6,850	0	6,850	0	0	0	0	2,277	0	591	4	6,850	6,850
RO	2012	0	0	0	7,598	0	0	4	779	6,042	0	4	7,598	7,598
	2013	5	0	0	7,560	0	0	4	777	4,804	0	4	7,560	7,560
	2014	0	0	0	7,508	0	0	4	777	4,785	0	4	7,508	7,508
RS	2012													
	2013	6501	0	0	0	0	0	0	139	6501	6501	4	6501	6501
	2014													
SE	2012	0	6,628	6,628	1,748	0	0	3	0	0	2,608	4	6,628	6,628
	2013	0	6,201	6,201	3,137	0	0	3	0	0	2,226	4	6,201	6,201
	2014	0	5,800	5,800	4,423	0	0	3	0	0	1,377	4	5,800	5,800
SI	2012	9,205	0	0	0	0	0	0	2,749	7	773	4	9,205	9,205
	2013	9,001	0	0	0	0	0	0	2,767	7	772	4	9,001	9,001
	2014	0	0	0	0	0	0	1	2,811	7	774	4	9,189	9,189
SK	2012	0	5,291	5,291	5,291	0	0	1	0	0	0	4	5,291	5,291
	2013	0	5,402	5,402	5,402	0	0	1	0	0	0	4	5,402	5,402
	2014	0	5,490	5,490	5,490	0	0	1	0	0	0	4	5,490	5,490
UK	2012	0	0	10,175	961	0	0	12	709	0	98	1	10,175	10,175
	2013	0	0	10,172	979	0	0	12	710	0	92	0	10,172	10,172
	2014	0	0	9,860	959	0	0	12	710	0	98	1	9,860	9,860

Note: (1)DB040 identifies the Region for each observation; DB060 identifies the Primary sampling unit (PSU); DB062 identifies the Secondary sampling unit; DB070 identifies the order of selection of PSU; DB075 identifies the rotation group; and DB030 identifies the household ID.
(2)The number of observations refers to the number of households in the data.
Source: EU-SILC 2012 UDB, version 4; EU-SILC 2013 UDB (version 3); EU-SILC 2014 UDB (version 1).

Table 4: Sample design variables constructed for EU-SILC 2010-2014 (D-File)

Country	Strata					PSU					Total observations				
	2010	2011	2012	2013	2014	2010	2011	2012	2013	2014	2010	2011	2012	2013	2014
AT	1	1	1	1	1	6,188	6,187	6,232	5,977	5,909	6,188	6,187	6,232	5,977	5,909
BE	3	3	3	3	3	274	274	274	275	275	6,132	5,910	5,817	6,159	6,021
BG	2	1	1	1	1	1,856	6,554	5,706	4,971	4,963	6,171	6,554	5,706	4,971	4,963
CH	1	1	1	1		7,513	7,502	7,529	7,341		7,513	7,502	7,529	7,341	
CY	1	1	1	1	1	3,780	3,917	4,638	4,648	4,294	3,780	3,917	4,638	4,648	4,294
CZ	8	8	8	8	8	1,671	1,721	1,661	1,589	1,551	9,098	8,866	8,773	8,275	8,053
DE	1	1	1	1	1	13,079	13,512	13,145	12,703	12,744	13,079	13,512	13,145	12,703	12,744
DK	1	1	1	1	1	5,867	5,331	5,355	5,419	5,758	5,867	5,331	5,355	5,419	5,758
EE	1	1	1	1	1	4,972	4,993	5,433	5,775	5,871	4,972	4,993	5,433	5,775	5,871
EL	1	4	4	4	4	1,170	1,120	1,094	1,340	1,536	7,005	6,029	5,626	7,439	8,620
ES	18	18	18	18	18	2,000	1,997	1,996	1,989	1,989	13,597	13,109	12,714	12,139	11,965
FI	1	1	1	1	1	10,989	9,351	10,307	11,370	11,030	10,989	9,351	10,307	11,370	11,030
FR	75	22	22	22	22	4,106	1,197	1,205	11,090	1,049	11,044	11,360	11,999	11,131	11,384
HR	1	1	1	1	1	644	1,262	1,373	1,470	1,567	3,703	6,403	5,853	5,362	5,443
HU	1	1	1	1	1	5,531	6,530	6,169	5,598	4,996	9,813	11,685	11,311	10,223	9,211
IE	1	1	1	1	1	2,462	1,998	1,243	1,150	1,526	4,642	4,333	4,592	4,922	5,486
IS	1	1	1	1	1	3,021	3,018	3,091	3,020	3,001	3,021	3,018	3,091	3,020	3,001
IT	111	106	113	111	115	6,497	6,978	7,448	6,829	7,362	19,147	19,399	19,579	18,487	19,663
LT	1	1	1	1	1	5,314	5,200	5,394	5,142	5,194	5,314	5,200	5,394	5,142	5,194
LU	1	1	1	1	1	4,876	5,464	6,031	3,770	3,879	4,876	5,464	6,031	3,770	3,879
LV	1	1	1	1	1	1,156	1,255	1,177	1,104	1,120	6,255	6,599	6,499	6,309	6,125
MT	1	1	1	1	1	3,781	4,076	4,350	4,381	4,381	3,781	4,076	4,350	4,381	4,381
NL	1	1	1	1	1	444	448	439	435	10,174	10,134	10,492	10,168	10,131	10,174
NO	1	1	1	1	1	5,227	4,628	6,050	6,031	7,371	5,227	4,628	6,050	6,031	7,371
PL	6	1	1	1	1	449	12,871	13,116	12,899	12,978	12,930	12,871	13,116	12,899	12,978
PT	1	1	1	1	1	541	542	542	1,994	2,277	5,182	5,740	6,257	6,491	6,850
RO	4	4	4	4	4	778	780	779	777	777	7,718	7,675	7,598	7,560	7,508
RS				1					139					6,501	
SE	1	1	1	1	1	7,173	6,717	6,628	6,201	5,800	7,173	6,717	6,628	6,201	5,800
SI	1	1	1	1	1	2,725	2,761	2,749	2,767	2,811	9,364	9,247	9,205	9,001	9,189
SK	1	1	1	1	1	5,376	5,200	5,291	5,402	5,490	5,376	5,200	5,291	5,402	5,490
UK	2	2	2	2	2	1,181	1,155	1,669	1,688	1,668	8,109	8,058	10,175	10,172	9,860

Source: EU-SILC UDB 2010 (version 6), 2011 (version 4), 2012 (version 4), 2013 (version 3) and 2014 (version 1)

Box 1 describes the changes in the syntax for 2012, 2013 and 2014 when compared to 2011. The update of EU-SILC dataset from 2011 to 2012, 2013 and 2014 did not require major adjustments in the syntax.

Box 1

Czech Republic

Up to 2011, DB060 was not unique across panels. As a result, PSUs were identified on the basis of a combination DB060 and DB075. Since 2012, DB060 is unique across panels.

France

In the 2011 EU-SILC for France, elements of panel 6 received the same PSU code. Because of that, the country required individual treatment when identifying PSUs, setting the variable *psutest* to missing value when the DB075 was equal to 6. In 2012 and 2013 this has been corrected. As previously noted, DB060 is largely missing for 2013, but not for 2012 nor 2014. Therefore, for 2013, the constructed PSU variable is mainly based on household IDs instead of DB060.

Italy

For the 2014 version 1 release the coding of the flag variable for DB060 has changed in line with the suggestions from Goedemé (2013b). The flag DB060_F==2 identifies which PSUs remain in the sample for the entire duration of the EU-SILC. In this case, an adjustment in the codes is required and the condition "if npanels>=2" must be replaced by "DB060_==2".

United Kingdom

As previously noted, prior to 2012, all households for cross-sectional and longitudinal EU-SILC originated from the General Lifestyle Survey (GLF) sample. Since 2012, cross-sectional and wave 1 respondents from the longitudinal panel have been selected from the Family Resources Survey (FRS). When describing the sample size for the 2012 Cross-sectional data, UK's Quality Report 2012 only provides information about the 1st rotational group, which is composed by the entire sample. This explains the fact that DB075 has only one value (Table 3), which required changes in the syntax for 2012 and 2013. Before 2012, the self-representing PSU for Northern Ireland was identified according to three conditions: the PSU had to appear in all 4 panels; PSU had to have the largest weighted number of households; and DB070 should not be filled. From 2012 onwards, the test had to be modified once the country's cross-sectional sample only provides information for one rotational group (DB075) due to the changes in the survey instrument. Now the identification of Northern Ireland is considerably simplified, due to the availability of DB040. In addition, until 2011, if households had moved to another postcode sector, they would be represented by a new DB060 code. This has been corrected for the 2012, 2013 and 2014 versions.

Republic of Serbia

EU-SILC 2013 provides for the first time information on the Republic of Serbia. This addition did not require any changes in the syntax and the excel output tables were automatically adjusted to include the new country. The country's quality report has not yet been released and there is no information available about its sample design. Therefore, a simple random sample of PSUs is assumed, which gives the most conservative sampling variance estimates. Alternatively, one could assume a simple random sample of households, which would be less conservative.

4 Using the sample design variables in practice

The Stata syntax for the construction of sample design variables is applied to the cross-sectional EU-SILC 2012 UDB (version 4), EU-SILC 2013 UDB (version 3) and EU-SILC 2014 UDB (version 1). Syntaxes for previous years are available in the format of Stata do-files at http://timgoedeme.com/eu-silc-standard-errors/. In Annex A below, we include a commented syntax as an example. Along with the syntaxes, csv files containing the constructed sample design variables strata1 and psu1 are also available. These files must be converted to Stata dta files using the insheet command (or to the proper format for other statistical software packages). After that, they must be merged one-to-one to the original EU-SILC D-File, using DB010, DB020, and DB030 as the merging variable list.When using these syntaxes, please refer to Goedemé (2013a) as well as this note. The syntax file must be executed using the EU-SILC D-file. After this, the EU-SILC D-file will be ready to be merged to additional EU-SILC files. Before explaining the Stata do-files, we briefly highlight how the new sample design variables should be used.

Stata command *svyset* allows the use of the sample design variables when calculating point estimates based on complex survey data. In order to estimate standard errors considering the sample design variables *strata1* and *psu1*, Stata requires them to be declared while setting the dataset to survey design using the command *svyset*. This command declares the data to be complex survey data and designates variables that contain information about the sample design. It must be used before using any *svy* command (Judkins, 1990). Similarly, in SPSS *CSPLAN* must be specified and the commands for complex sample data must be used for further data analysis (e.g. *CSDESCRIPTIVES*). In SAS it works somewhat differently, in that sample design variables can be specified for each command that is used (e.g. *PROC SURVEYFREQ*). The following syntax exemplifies the estimation of two indicators in Stata using svyset and svy: at risk of poverty (*arop60*)[3] and material deprivation (*dep4*)[4], by country.

```
<<insert code for loading the file with indicators already
constructed>>

        . svyset psu1 [pw=rb050], strata(strata1)
        . svy: tab country arop60, row per se
        . svy: tab country dep4, row per se
```

The effect of the use of sample design variables on standard errors is illustrated in Table 5. The table depicts point estimates for both indicators along with standard errors (SE) type (1) and (2). SE (1) refers to standard errors that have been calculated without considering the sample design variables, while SE (2) refers to standard errors that have been calculated considering the sample design variables. Standard errors take account of the fact that the poverty line has been estimated on the basis of the data using the DASP module developed for Stata (Araar and Duclos, 2007). For all point estimates, SE values are larger when the sample design variables are considered. The greater the standard errors the wider will be the confidence intervals. As previously discussed, when a confidence interval is wider the uncertainty about the point estimate to which it refers is greater and should be interpreted more cautiously.

[3] Being at-risk-of-poverty means living in a household with an equalized net disposable household income below 60 percent of the national median (Goedemé, 2013a)

[4] Severe material deprivation is measured by an index of nine items relating to financial stress and the enforced lack of some durables. All persons living in a household which at the moment of the interview lacks at least 4 out of 9 items are considered severely materially deprived (Goedemé, 2013a)

Table 5: Sample design variables effect on standard errors for at risk of poverty (AROP60) and material deprivation, EU-SILC 2012

Country	At risk of poverty (AROP60)			Material deprivation		
	P.E. (%)	Standard error		P.E. (%)	Standard error	
		(1)	(2)		(1)	(2)
AT	14.4	0.325	0.596	4	0.195	0.349
BE	15.3	0.342	0.581	6.3	0.265	0.547
BG	21.2	0.326	0.643	44.1	0.467	0.912
CH	15.9	0.306	0.562	0.8	0.099	0.18
CY	14.7	0.343	0.639	15	0.358	0.705
CZ	9.6	0.236	0.441	6.6	0.205	0.411
DE	16.1	0.237	0.368	4.9	0.158	0.229
DK	13.1	0.514	0.702	2.8	0.305	0.446
EE	17.5	0.395	0.612	9.4	0.313	0.476
EL	23.1	0.415	0.713	19.5	0.513	1.054
ES	20.8	0.262	0.473	5.8	0.191	0.372
FI	13.2	0.284	0.42	2.9	0.154	0.214
FR	14.1	0.235	0.603	5.3	0.164	0.32
HR	21	0.329	0.744	15.9	0.346	0.923
HU	14	0.212	0.558	25.7	0.29	0.976
IE	15.7	0.41	0.754	9.8	0.347	0.687
IS	7.9	0.352	0.562	2.4	0.213	0.315
IT	19.5	0.216	0.399	14.5	0.218	0.59
LT	18.6	0.501	0.874	19.9	0.517	0.903
LU	15.1	0.437	0.82	1.3	0.15	0.28
LV	19.2	0.329	0.621	25.7	0.383	0.749
MT	15.1	0.38	0.719	9.2	0.329	0.642
NL	10.1	0.334	0.739	2.3	0.182	0.415
NO	10	0.29	0.41	1.7	0.145	0.2
PL	17.1	0.218	0.444	13.5	0.212	0.42
PT	17.9	0.316	0.608	8.6	0.252	0.584
RO	22.7	0.317	0.734	30.1	0.42	1.168
SE	14.2	0.303	0.456	1.3	0.105	0.15
SI	13.6	0.255	0.393	6.6	0.201	0.327
SK	13.2	0.289	0.574	10.5	0.269	0.546
UK	16	0.274	0.507	7.8	0.207	0.411

Note: (1) ”P.E.” refers to Point Estimates. (2) Standard errors in parenthesis. (3) S.E. (1) refers to standard errors that have been calculated considering the sample design variables, while S.E. (2) refers to standard errors that have been calculated without considering the sample design variables.
Source: Estimations based on EU-SILC UDB 2012 (version 4).

5 Final remarks

The aim of this note aims was to explain the update of EU-SILC UDB sample design variables for 2012, 2013 and 2014, based on the methodology developed by Goedemé (2010b, 2013a). Although several of the issues confronted while reconstructing the EU-SILC sample design variables are identical for all releases of the data, the update of EU-SILC dataset from 2011 to 2012, 2013 and 2014 required minor adjustments in some of the strategies used to compute the new sample design variables *psu1* and *strata1*. In particular, for the Czech Republic, France, the United Kingdom and the Republic of Serbia.

The effect of the use of the new sample design variables on standard errors is illustrated for ‘At risk of poverty’ and ‘Material deprivation’ and shows that neglecting the sample design can lead to an underestimation of standard errors. The results confirm the findings of previous studies (Rodgers and Rodgers,1993; Howes and Lanjouw, 1998; Jolliffe et al. 2004; and Goedemé, 2010b, 2013a, 2013b) and ascertains that indicators based on EU-SILC should be accompanied by appropriate estimates of their precision and statistical reliability.

References

[1] Araar, A., Duclos, J.Y. (2007). *DASP: Distributive Analysis Stata Package*. PEP, CIRPÉE and World Bank, Université Laval.

[2] Berger, Y., Osier, G. and Goedemé, T. (forthcoming). Standard error estimation and related sampling issues. In Atkinson, A.B., Guio,A.-C., and Marlier, E. (eds.) *Monitoring social Europe.* Luxembourg: Eurostat.

[3] Eurostat (2013). *2010 Comparative EU Final Quality Report* Luxembourg: European Commission, 25p.

[4] Goedemé, T. (2010a). *The construction and use of sample design variables in EU-SILC. A user's perspective*. Report prepared for Eurostat, November 2010, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp, 16p.

[5] Goedemé, T. (2010b). The standard error of estimates based on EU-SILC. An exploration through the Europe 2020 poverty indicators, *CSB Working Paper Series WP 10/09*, Antwerp, Herman Deleeck Centre for Social Policy, University of Antwerp.

[6] Goedemé, T. (2013a). How much Confidence can we have in EU-SILC? Complex Sample Designs and the Standard Error of the Europe 2020 Poverty Indicators in *Social Indicators Research*, 110(1): 89-110, doi:10.1007/s11205-011-9918-2.

[7] Goedemé, T. (2013b). The EU-SILC sample design variables: critical review and recommendations, *CSB Working Paper Series WP 13/02*, Antwerp: Herman Deleeck Centre for Social Policy.

[8] Heeringa, S. G., West, B. T., and Berglund, P. A. (2010). *Applied Survey Data Analysis*, Boca Raton: Chapman and Hall/CRC, 467p.

[9] Howes, S. and Lanjouw, J. O. (1998). Does Sample Design Matter for Poverty Rate Comparisons? in *Review of Income and Wealth*, 44(1): 99-109.

[10] Jolliffe, D., Datt, G. and Sharma, M. (2004). Robust Poverty and Inequality Measurement in Egypt: Correcting for Spatial-price Variation and Sample Design Effects in *Review of Development Economics*, 8(4): 557-572.

[11] Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6: 223-239.

[12] Kalton, G. (1979). Ultimate Cluster Sampling, in *Journal of the Royal Statistical Society, Series A (General)*, 142(2): 210-222.

[13] Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

[14] Osier, G. (2012). *The linearization approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards?*, Paper presented during the Workshop on standard error estimation and other related sampling issues in EU-SILC, organized in the context of the EU-funded "Net-SILC2" project, Eurostat, Luxembourg, 29-30 March 2012.

[15] Rodgers, J. R. and Rodgers, J. L. (1993). Chronic Poverty in the United States, in *The Journal of Human Resources*, 28(1): 25-54.

[16] Wolter, K. M. (2007). *Introduction to Variance Estimation*, New York: Springer, 447p.

A Computational construction of strata and PSU variables in Stata

```
*****************************************************************
*Re-construction EU-SILC sample design variables
*EU-SILC UDB 2012 (version 4)

*Author: Tim Goedemé, updated by Lorena Zardo Trindade

<<insert code for loading the D-file>>
```

Note that the variables in this syntax are in uppercase. The following command ensures that your dataset contains variable names in uppercase.

```
. foreach var of varlist _all {
        local newname = upper("`var'")
        cap rename `var' `newname'
        }
```

The construction of the new EU-SILC sample design variables requires a previous preparation of the D-file. First, the variables DB020 and DB030 are renamed as 'country' and 'hid', respectively, and country labels are stored in the global 'countries '.

```
. cap rename DB020 country
. cap rename COUNTRY country

. cap drop countryNR
. encode country, gen(countryNR)

. cap rename DB030 hid
. cap rename HID hid

. local varlist country
. sort `varlist'
. tempvar tesje
. qui: gen `tesje'=1 if `varlist'[_n]!=`varlist'[_n-1]
. sort `tesje' `varlist'
. qui: count if `tesje'==1
. local nrvalues=r(N)
. global countries
. local counter=1
. while `counter'<=`nrvalues' {
        local value1=`varlist'[`counter']
        local value2=`varlist'[`counter'-1]
        if "`value1'"!="`value2'" {
                global countries ${countries} `value1'
        }
        local counter=`counter'+1
        }
. global ncountries=wordcount("${countries}")
. display "${countries}"
. display "number of countries in datafile: " $ncountries
```

After the initial preparation, PSUs are identified by DB060 in a new variable *psutest*. However, this identification should not be applied to Poland, Bulgaria, Czech Republic, Italy and United Kingdom. These countries require individual treatment when identifying PSUs.

```
. cap drop psutest
. gen double psutest=DB060
```

Poland and Bulgaria: similar cases in which the variable DB060 is not unique across strata. Consequently, *psutest* is replaced by missing values for these countries and the identification of PSUs will be done in a later stage of the syntax.

```
. replace psutest=. if country=="BG" | country=="PL"
```

Italy: the dataset refers to a two stage sample design survey with rotation at PSU level, suggesting that large municipalities are self-representing and might remain in the sample across different panels. Therefore, for those PSUs considered as self-representing, *psutes*t must be replaced by missing values. The following commands replaces *psutest* by missing values for those PSUs (DB060) that appear in at least three out of four panels (DB075). The identification of PSUs will be done in a later stage of the syntax. . For the 2014 version 1 release the coding of the flag variable for DB060 has changed in line with the suggestions from Goedemé (2013b) and it indicates when the PSU remains in the sample for the entire duration of the EU-SILC. In this case, *psutest* will be replaced by missing values for those PSUs (DB060) that have being coded DB060_F==2.

```
. cap drop tester
. cap drop npanels

. sort country DB060 DB075
. gen tester=.
. replace tester=1 if DB060[_n]==DB060[_n-1] & DB075[_n]!=DB075[_n-1]

. bysort country DB060: egen npanels=sum(tester)

. sort country DB060
. ta npanels if country=="IT" & DB060[_n]!=DB060[_n-1]
. ta npanels if country=="IT"
. replace psutest=. if npanels>=2 & country=="IT"
```

The following commands generate the variable *groupsit* that will be used in a later stage of this syntax, when constructing the stratum variable. First, a variable *tester* is defined by DB060 for the Italian PSUs that appear in at least three out of four panels. Then, the *groupsit* variable is generated by the sorted ascending order of *tester*.

```
. gen tester=DB060 if npanels>=2 & country=="IT"
. cap drop groupsit
. gsort tester, gen(groupsit)
```

United Kingdom: Northern Ireland (DB040=UKN) is a self-representing PSU and can be identified as the unit with the largest number of households.[5] As in the case of Italy, the PSU identification must be replaced by missing values in *psutest* and with missing values for DB070 when DB040 is not available. Since DB040 is available, the original PSU identification can be replaced by missing values in *psutest* and the identification of PSUs will also be done in a later stage of the syntax..

```
. cap drop cons
. gen cons=1 if country=="UK"
. cap drop nrpsu
. bysort country DB060: egen nrpsu=total(cons==1) if country=="UK"

. sum nrpsu if DB040=="UKN"
. local max=r(max)
. replace psutest=. if nrpsu==`max' & country=="UK"
```

[5] Self-representing PSU is itself a stratum, and PSUs within this stratum are households.

The individual treatments for PSU identification is followed by the construction of the sample design variables. The stratum variable is the first to be defined. As the original stratification variable (DB050) is lacking, the variable that identifies NUTS1 / NUTS2 regions (DB040) can be used as proxy. However, in many countries this variable seems to be inconsistent as it underestimates the number of strata. Its use must be done with caution and only for the following countries: Belgium, Czech Republic, Greece, Spain, France, Italy and Romania.

For these countries, the first step in constructing the stratum variable is to create *region0* based on DB040. In the case of Spain, the regions regarding Ceuta (ES63) and Melilla (ES64) must be grouped as ES80 since they are part of the same stratum. Then, *region1* is created based on *region0.* For the countries that DB040 should not be used as stratum, *region1* is set to 0. Different treatment must be applied to regions with self-representing PSUs. For the Italian self-representing PSUs, *region1* is replaced by the sum of *groupsit* and the total number of regions in *region1*, 61.

```
. global stratcs BE CZ EL ES FR IT RO
. cap drop region0
. gen region0=""
. foreach ctry of global stratcs {
        replace region0=DB040 if country==`"`ctry'"'
        }
. replace region0="ES80" if DB040=="ES63"|DB040=="ES64"

. cap drop region1
. encode region0, gen(region1)
. replace region1=0 if region1==.
. sum region1
. local min=r(max)
. replace region1=groupsit+`min' if country=="IT" & npanels>=2
```

After the previous step, the stratum variable *strata0* is defined by a combination of *region1* and the numeric country code in *countryNR* (multiplied with a multiple of 10 such that all strata are unique, and start with the country code number).

```
. sum region1
. local minimum=r(max)
. local maximum=10

. while `maximum'<=`minimum' {
        local maximum=`maximum'*10
  }
. cap drop strata0
. gen strata0=countryNR*`maximum'+region1

. sum strata0 if country=="UK"
. local stratum=r(max)+2
. replace strata0=`stratum' if country=="UK" & psutest==.

. sum strata0
```

The next step in constructing the sample design variables is to define the PSU variable. The new PSU variable is defined as the stratum code generated earlier, followed by the PSU code (*psutest*), or – if missing - the household identifier. In order to make sure that all PSU codes are unique, we first multiply the stratum code with a multiple of 10 such that when household ID is added, the codes remain unique across countries.

```
. sum hid
. local minimum1=r(max)
. local maximum1=10
```

```
. while `maximum1'<=`minimum1' {
      local maximum1=`maximum1'*10
  }

. sum psutest
. local minimum2=r(max)
. local maximum2=10

. while `maximum2'<=`minimum2' {
      local maximum2=`maximum2'*10
  }

. cap drop psu0
. gen double psu0=.
. replace psu0=strata0*`maximum2'+hid/`maximum1'
. replace psu0=strata0*`maximum2'+psutest if psutest!=.

. sum psu0
```

In the case of several countries, stratification by DB040 causes PSUs to be split across regions because of households moving between the moment of selection and the moment of interview. Hence, households that have moved should be reallocated to the correct stratum. In the countries for which PSU (DB060) is not a missing value, no households have moved between the moment of selection and the moment of interview. The first step is to regroup split households creating two variables *nocheck* and *checker*. The variable *nocheck* identifies as 1, all observations belonging to the Italian self-representing PSUs and the observations for which *psutest* has missing values. These observations do not need to be checked for split PSUS. Then, after sorting the data by *country*, *psutest* and *hid*, *checker* is created as a numeric variable that assumes value 1 if the observation belongs to a split PSU (if *psu0* for observation *n* is different than observation *n-1* and *psutest* for observation *n* is equal to observation *n-1* and *nocheck* is different than 1). In 2012, PSUs from Belgium, Czech Republic, Spain, France, Italy and Romania have been split by the stratification procedure, therefore must be regrouped. PSUs from Greece have not been split.

```
. global countrypsu BE CZ ES FR IT RO
. cap drop checker
. gen checker=.

. sort country psutest hid

. cap drop nocheck
. gen nocheck=1 if psutest==. | ((npanels>=2) & country=="IT")

. replace checker=0 if psu0[_n-1]!=psu0[_n] & ///
 psutest[_n-1]!=psutest[_n] | nocheck==1
. replace checker=0 if psu0[_n-1]==psu0[_n] & ///
 psutest[_n-1]==psutest[_n] & nocheck!=1
. replace checker=1 if psu0[_n-1]!=psu0[_n] & ///
 psutest[_n-1]==psutest[_n] & nocheck!=1
. replace checker=2 if psu0[_n-1]==psu0[_n] & ///
 psutest[_n-1]!=psutest[_n] & nocheck!=1

. ta country checker
```

Occasionally, when DB060 is not unique across DB040 and the latter variable is a poor substitute for the real stratum (DB050), PSUs might be split on purpose. In cases like this, *checker* must be reset to 0 and PSUs should not be regrouped (e.g. Poland and Bulgaria until 2010). For the 2012 and 2013 releases, this procedure has not been applied to any country, for reasons explained above.

```
. foreach ctry of global countries {

        di "`ctry'", _continue
        replace checker=0 if country=="`ctry'" & ///
        strpos("${countrypsu}", "`ctry'")==0
  }
. sort country psu0
. foreach ctry of global countrypsu {
        tab country checker if country=="`ctry'" & psu0[_n]!=psu0[_n-1]
  }
```

Now, once the split PSUs have been identified, strata1 is created as a replica of *strata0*, adjusted by the reallocation of split PSUs to the correct stratum (that is, the stratum with the highest number of observations for each split PSU). In some cases, the number of households of the same PSU is equally distributed across two or more strata. In these cases, the 'correct' stratum is even more difficult to guess, and we take the one with the lowest stratum number.

```
cap drop strata1
gen strata1=strata0

. foreach ctry of global countrypsu {
        global psu`ctry'
        di "`ctry'"
        tab psutest if country=="`ctry'" & checker==1, matrow(psu`ctry')
        local rows=rowsof(psu`ctry')
        forvalues x=1/`rows' {
                local nr=el(psu`ctry', `x',1)
                global psu`ctry' ${psu`ctry'} `nr'
        }
        di "${psu`ctry'}"
  }

. foreach ctry of global countrypsu {
        di "`ctry'"

        foreach psu of global psu`ctry' {
                local check1
                local check2
                local check3

                tab psutest strata0 if country=="`ctry'" & \\\
                psutest==`psu', matcell(freq1) matcol(stratname
                local cols=r(c)
                forvalues y=1/`cols' {
                local check1=el(freq1, 1, `y')
                if `y'<`cols' {
                        local check2 `check2' `check1',
                }
                if `y'==`cols' {
                        local check2 `check2' `check1'
                }
        }
        local check3=max(`check2')

        forvalues y=1/`cols' {
                if el(freq1, 1, `y')==`check3' {
                        replace strata1=el(stratname, 1, `y') if \\\
                        (country=="`ctry'" & psutest==`psu')
```

```
                        di "`ctry' `psu': "el(stratname, 1, `y')
                        continue, break
                        }
                }
        }
}
```

Since *psu0* is constructed based on *strata0*, the variable also must be adjusted to consider the reallocation of split PSUs to the correct strata. Therefore, *psu1* is created as a duplicate of *psu0* for all observations with missing values for *psutest*. However, when *psutest* is not a missing value, *psu1* is replaced by the *strata1* codes by a maximum value, which allows PSU codes to be added to *strata1* codes without overwriting them.

```
. qui: sum psutest
. local minimum2=r(max)
. local maximum2=10
. while `maximum2'<=`minimum2' {
        local maximum2=`maximum2'*10
  }
. cap drop psu1
. gen double psu1=psu0
. replace psu1=strata1*`maximum2'+psutest if psutest!=.
```

The sample design variables *strata1* and *psu1* are the main outputs regarding the construction of EU-SILC sample design variables. However, the following step allows the user to produce user-friendly excel tables containing the total number of strata, PSUs and observations for each country in the EU-SILC UDB. The missing values for strata and PSUs are also reported by country. In addition, variables that are not required for further analyses are dropped, and the variables *country* and *hid* are renamed according to their original nomenclature *DB020* and *DB030*, respectively.

```
. drop countryNR psutest groups cons nrpsu npanels tester groupsit \\\
region0 region1 strata0 psu0 checker nocheck
. cap drop nhid
. local vals 1
. foreach x of local vals {
        svyset psu`x' [pw=DB090], strata(strata`x')

        cap mat drop svy`x'
        preserve
        foreach ctry of global countries {
                cap restore, preserve
                di "********************"
                di "`ctry'"
                di "********************"

                keep if country=="`ctry'"

                cap drop single`ctry'
                svydes if country=="`ctry'"
                local nsingle=r(N_single)
                local misstrat=r(N_mstrata)
                local mispsu=r(N_munits)
                local misobs=r(N_miss)
                local nstrats=r(N_strata)
                local npsu=r(N_units)
                local nobs=r(N)
                mat svy`x'=(nullmat(svy`x') \ `nsingle', ///
                `misstrat',  `mispsu', `misobs', `nstrats', ///
```

```
                        `npsu', `nobs')
                        cap drop single`ctry'
                        }
                restore
                mat rownames svy`x'=${countries}
                mat colnames svy`x'=nsingle misstrat mispsu misobs ///
                nstrats npsu nobs
                mat li svy`x'

                xml_tab svy`x', save("<< insert directory >>\ ///
                << insert name of excel file >>") newappend sheet(svy`x')
          }

        << xml_tab is a user-written command that can be downloaded for free >>

        . cap rename country DB020
        . cap rename hid DB030

        . compress
        . save "<< insert directory >>\2012-d2_sdv.dta", replace

end of do-file
```

B List of EU-SILC countries

AT Austria
BE Belgium
BG Bulgaria
CH Switzerland
CY Cyprus
CZ Czech Republic
DE Germany
DK Denmark
EE Estonia
EL Greece
ES Spain
FI Finland
FR France
HR Croatia
HU Hungary
IE Ireland
IS Iceland
IT Italy
LT Lithuania
LU Luxemburg
LV Latvia
MT Malta
NL Netherlands
NO Norway
PL Poland
PT Portugal
RO Romenia
RS Serbia
SE Sweden
SI Slovenia
SK Slovakia
UK United Kingdom