gesis Leibniz-Institut für Sozialwissenschaften

GESIS Papers

2018|16

EU-SILC Tools:

Calculating Standard Errors for EU-SILC using SPSS

Anika Herter & Heike Wirth

GESIS Papers 2018|**16**

EU-SILC Tools:

Calculating Standard Errors for EU-SILC using SPSS

Anika Herter & Heike Wirth

GESIS – Leibniz-Institut für Sozialwissenschaften 2018

GESIS Papers

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

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

1 Introduction

The German Microdata Lab[1] (GML) at GESIS Leibniz Institute for the Social Sciences offers research-based services for researchers working with microdata from European and German official statistics. Our services include

- tools for data management and data analysis,
- a Microdata Information System called MISSY[2], which offers comprehensive data documentation for selected official microdata, and
- knowledge transfer[3] (training, conferences) concerning official microdata.

This paper focuses on the calculation of the standard error for EU-SILC data using **SPSS (download: https://www.gesis.org/gml/european-microdata/eu-silc/).** It goes back to the work of Tim Goedemé and Laura Zardo Trindade, who developed **Stata-t**ools for the correct calculation of the standard error for EU-SILC. Also, Tim Goedemé makes CSV files (containing reconstructed strata and PSU variables) available for download on his homepage that can be merged with the Eurostat EU-SILC User Database (UDB).[4] Nevertheless, we were often asked by researchers to provide a corresponding tool for SPSS. Therefore, we have 'translated' the Stata-tool from Tim Goedemé for the EU-SILC surveys from 2010 to 2014 (cross-sectional) into an SPSS-tool.

In the following we briefly outline the different sample designs used in EU-SILC. Then, using the AROPE indicator as an example, we compare the standard errors and confidence intervals resulting from specifying a simple sample design versus specifying a complex sample design (i.e., as calculated by Eurostat). Even this plain comparison illustrates that the standard errors tend to be underestimated when a simple random design is assumed. In the next step we show the standard errors and confidence intervals of the AROPE indicator per country as reconstructed with SPSS using the information available in the User Database. Appendix A shows an example of SPSS syntax and Appendix B contains instructions for creating a sampling plan with SPSS (prerequisite: availability of SPSS Complex Sample).

[1] https://www.gesis.org/en/gml/gml-home/

[2] http://www.gesis.org/missy/

[3] https://www.gesis.org/angebot/veranstaltungen/veranstaltungsarchiv/german-microdata-lab/

[4] https://timgoedeme.com/eu-silc-standard-errors/

2 Simple Random Sample versus Complex Sample Design

2.1 Sample Designs used in EU-SILC

The European Statistics on Income and Living Conditions (EU-SILC) is the main source for comparative statistics on income distribution, social exclusion, living conditions and poverty in the European Union. The rich data source is used to monitor social protection and social inclusion processes. That is, many indicators for evaluating the EU 2020 strategy of the European Union, such as the people at-risk-of-poverty or social exclusion rate, are generated using EU-SILC data (Atkinson, Guio, & Marlier, 2017).

EU-SILC data is collected by national statistical institutes (NSIs) of 35 countries and brought together and provided by Eurostat. As pointed out by Verma and Betti (2009) EU-SILC is a flexible instrument: It is made up of various data sources (registers and surveys); the data are collected at household and person level; both cross-sectional and longitudinal information is available. Similar flexibility can be found in the sample design of EU-SILC, which varies across countries. The sample designs used in the respective countries are documented in the National Quality Reports.[5] A summary of the national sample designs can be found in the Comparative Quality Reports provided by Eurostat (Eurostat 2013: Annex 3). Both the national and the Eurostat quality reports are publicly available on the Eurostat website.

Table 1 shows simplified the sampling designs used for EU-SILC in the different countries for 2013.[6] As can be seen in the table, the spectrum of sample designs ranges from (non-stratified) simple random sampling to stratified multi-stage sampling. A total of five countries do not use any stratification: Sweden uses a systematic sample and Denmark, Malta, Iceland, and Norway use a simple random sample. All other countries use stratified sampling techniques. The majority of them use a multi-stage sample, except for Germany, Cyprus, Lithuania, Luxembourg, Austria, Slovakia, and Switzerland, who use a stratified simple random sampling design. Estonia uses a systematic, stratified sample. Hungary is the only country that applies a different sampling scheme for each rotation group. Depending on the country the sample units can be individuals, households, dwellings or addresses.

Except for Denmark, Malta, Iceland, and Norway, all countries use more or less complex sampling procedures involving stratification and clustering. Compared to simple random sampling, the standard errors in a complex sample design can be larger or (rather rarely) smaller. Accordingly, the design variables must be taken into account for data analysis. Otherwise, the statistical program assumes a simple random sampling design, which usually leads to an underestimation of the standard errors.

[5] See: http://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports.

[6] Please note that these sampling design specifications may change over time.

Table 1. EU-SILC Sampling design by country

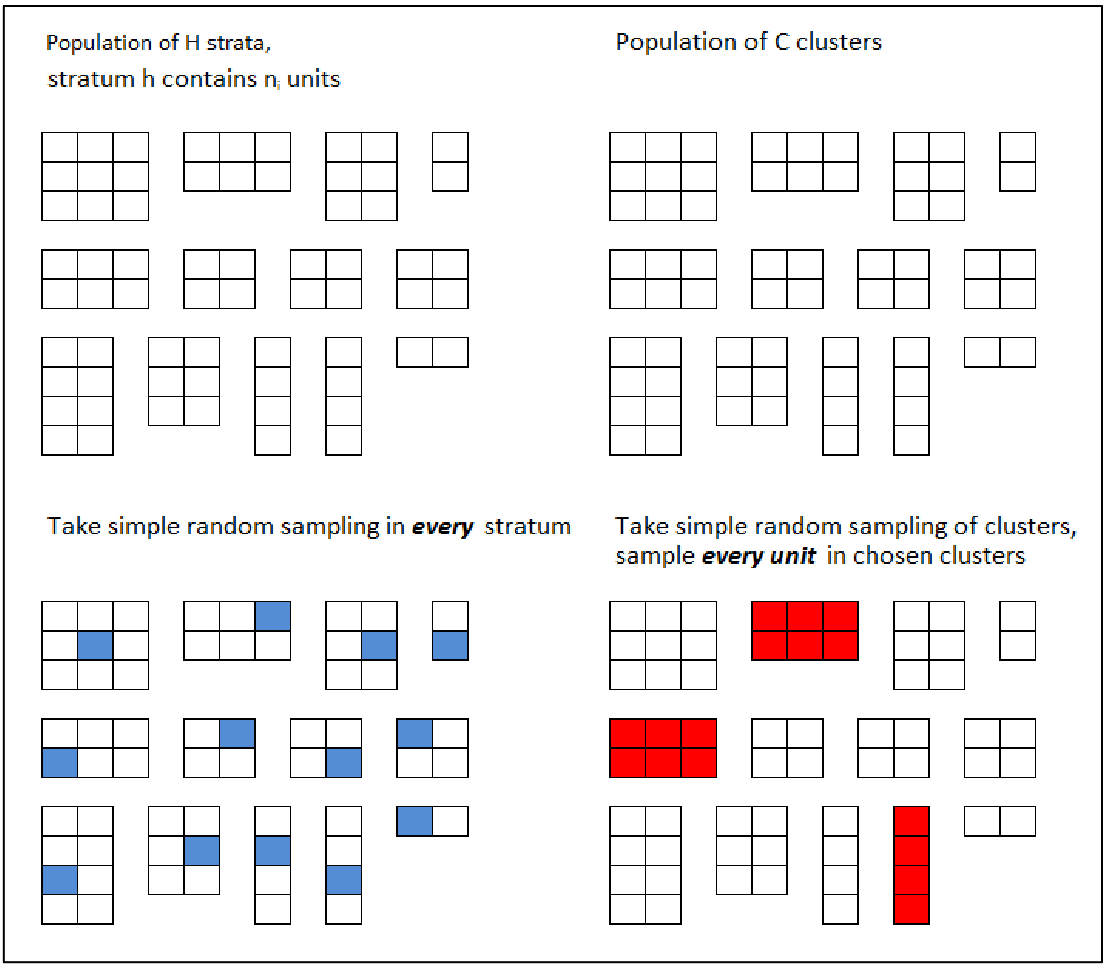
Sampling Design (in simplified terms)	Sampling Unit	Country	Variables used by Eurostat for variance estimation	
Without Stratification			**Strata**	**PSU**
Simple random sampling	Individuals	DK, IS, NO	-	DB030
	Dwellings/addresses	MT	-	DB030
Systematic sampling	Individuals	SE	-	DB030
With Stratification				
Stratified two-stage sampling	Dwellings/addresses	HR, LV, NL, PT	DB050	DB060
	Individuals	SI	DB050	DB060
Stratified multi-stage sampling	Dwellings/Addresses	CZ, EL, ES, FR, PL, RO, UK	DB050	DB060
	Households	BE, BG, IE, IT	DB050	DB060
Stratified simple random sample	Dwellings/addresses	AT, CY, DE*, LU	DB050	DB030
	Households	CH, SK	DB050	DB030
	Individuals	LT,	DB050	DB030
Stratified sampling according to a different design by rotational group	Households	HU	DB050	DB060
Stratified and systematic sampling	Individuals	EE	DB050	DB030
Stratified two-phase sampling	Individuals	FI	DB050	DB030

*From former participants of micro-census
Sources: Eurostat 2013, Eurostat 2016.
DB030: household ID; DB060: Primary Sampling Unit (PSU); DB050: Primary Strata

Stratification is a method to divide the population into subgroups (see Figure 0). Stratification is intended to ensure that the sample is representative regarding stratification criteria. Therefore, sampling is carried out within each subgroup or stratum (Babbie, 2013: 150). If there are differences in the variation of a characteristic across multiple strata, an estimate based on a stratified sample is generally more precise than an estimate based on a simple random sample. Thus, stratification increases the precision of a sample and reduces the standard error. Ignoring stratification can lead to an overestimation of standard errors (Schnell, Hill, & Esser, 2008: 279 ff.).

For *clustering*, the population is divided into clusters which can be for example municipalities (see Figure 0). From these clusters, a sample is drawn initially, then the units of each selected cluster are subsampled afterward (Babbie, 2013: 153). Assuming that there are greater similarities between units within one cluster as compared to independently selected units, clustering leads to a less uniform coverage of the population. The underestimation of the true population variance by clustering reduces the precision of estimates which is reflected in larger standard errors compared to a simple random sample (Howes & Lanjouw 1998).

Figure 0. Difference between Stratification and Clustering.


Population of H strata,
stratum h contains n_i units
Population of C clusters
Take simple random sampling in *every* stratum
Take simple random sampling of clusters, sample *every unit* in chosen clusters

Source: Based on Córdova/Zéphyr (2013)

In sum, the use of simple random sample formulas to calculate standard errors even though the data is based on a complex sample design is incorrect and can lead to misconclusions. The common statistical software assumes that the sample was carried out as a simple random sample. However, it is possible to modify these assumptions by defining the sample design variables. The relevant variables in EU-SILC – used by Eurostat to calculate standard errors - are the stratification variable (DB050), the primary sampling unit (DB060) and the household ID (DB030) (see Table 1). Unfortunately, the sampling design information is only partially included in the EU-SILC User Database (UDB) and some countries provide less sampling information than others.

2.2 Why the Sample Design matters - AROPE[7]-indicator

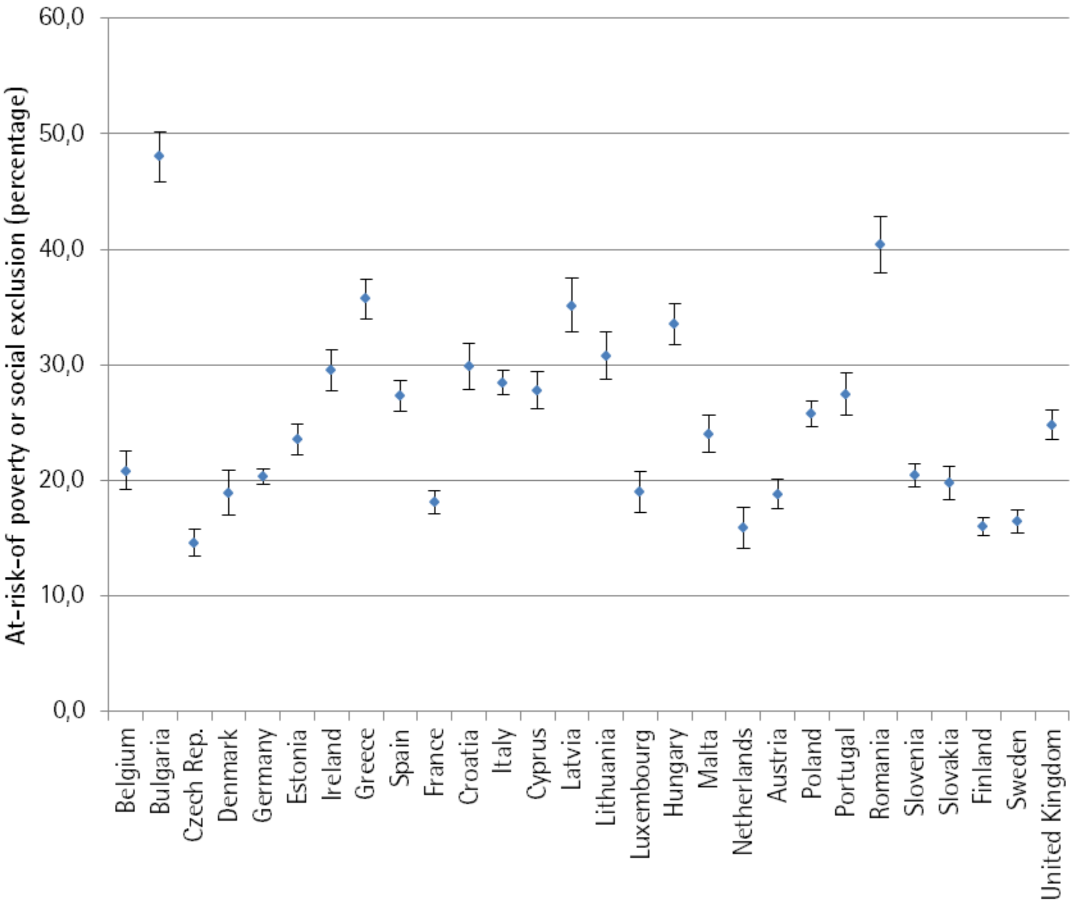
To illustrate why sample design matters when analyzing EU-SILC data, we use the AROPE indicator as an example. This indicator was developed within the framework of the EU 2020 reform strategy. As part of this strategy, the European Union has set specific targets, the achievement of which will be

[7] AROPE = People At-Risk-Of Poverty or social Exclusion

monitored by selected indicators. One headline indicator to monitor the poverty target is the *People at-risk-of poverty or social exclusion rate (AROPE).* It is the share of the population which is either at risk of poverty or severely materially deprived or lives in a household with very low work intensity. We focus on AROPE for reasons of comparison. For this indicator, Eurostat regularly publishes country-specific standard errors and confidence intervals. These are calculated using the sample design information as listed in Table 1, thus we could use them as a reference.

In the following, estimates of the AROPE-indicator by country are presented, whereby different sampling design information is used to calculate the standard error. Figure 1 shows the AROPE estimates computed by Eurostat considering the sample design information (DB030, DB050, and DB060, see Eurostat 2013: Appendix 4) of all countries.

Figure 1. AROPE-indicator (2013) and 95% confidence intervals as calculated by Eurostat using the proper design information.

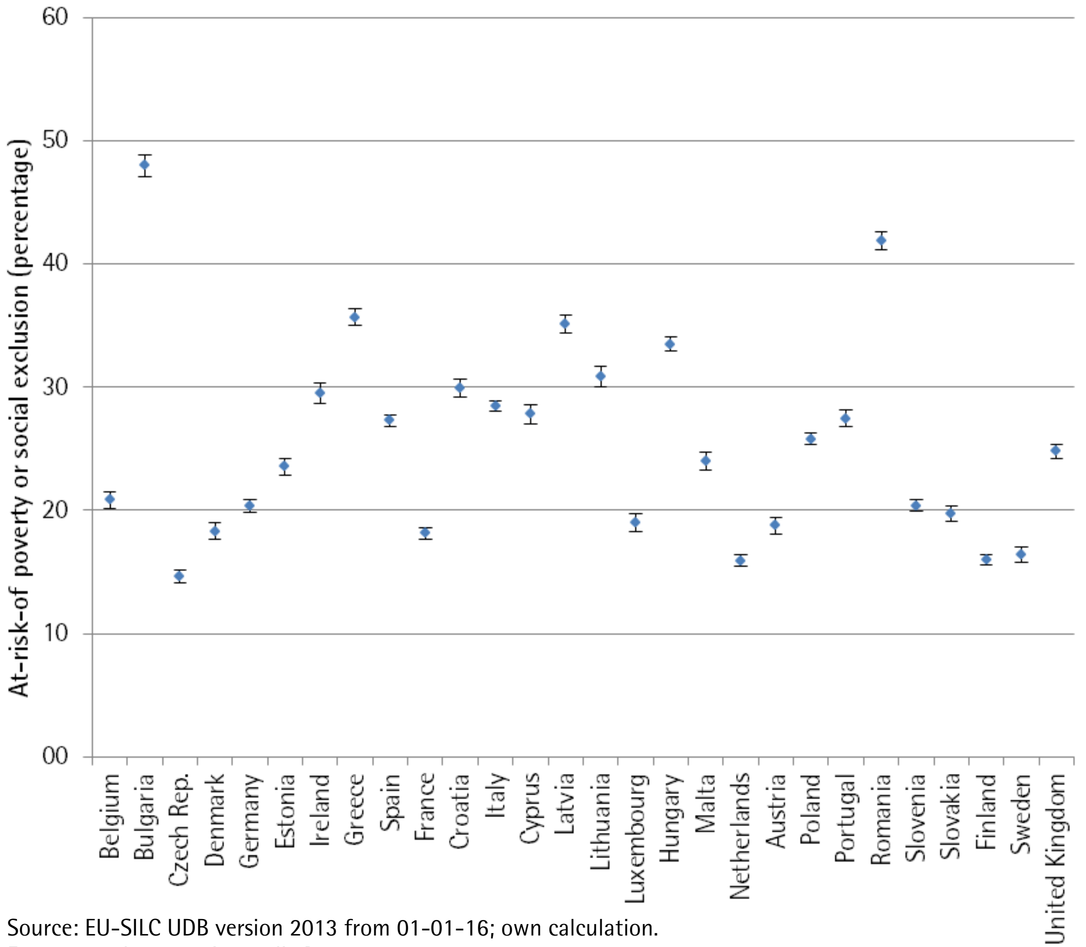


Source: Eurostat 2013, own illustration.
For exact values, see Appendix D.

Figure 2 displays the AROPE estimates, standard errors and confidence intervals calculated on the assumption that all countries use simple random sampling. Comparing both figures shows that the confidence intervals in Figure 1 are larger than in Figure 2. That implies that not taking the complex design information into account leads to an underestimation of the standard errors which in return could entail consequences for country comparison analyses. Compare for example Belgium and Denmark (see Figure 3 for illustrative purposes). In Belgium about 20.8 percent of the population is at risk of poverty and social exclusion, the respective share in Denmark is about 18.9 percent. These numbers are the same in Figures 1 and 2. However, looking at the confidence intervals, assuming a simple random selection, the difference between Belgium and Denmark is significant (as the respective confidence intervals do not overlap, see Figure 3). However, when the proper sample design information is taken into account (Figure 3), there is an overlap of the confidence intervals, i.e. the share of people at risk of poverty or exclusion in Belgium is not significantly different from Denmark. The same applies for other cross-country comparisons, for example when comparing France and Luxembourg.

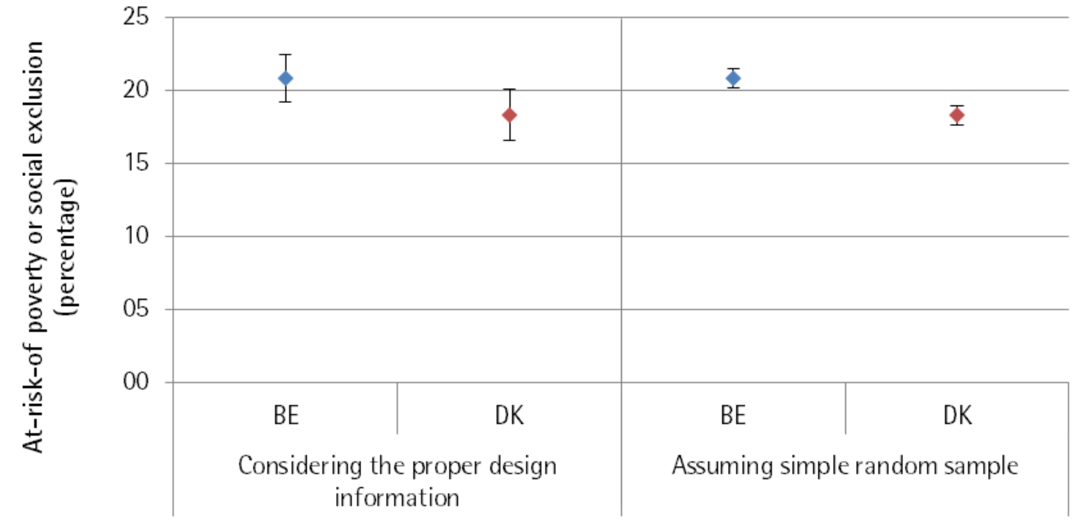
Thus, it is crucial to take country-specific sampling design information into account. However, the UDB of EU-SILC does not include all relevant information. The stratification variable (DB050) is not included at all, and the PSU variable (DB060) is partly missing or not properly coded for some countries (for a detailed discussion of these issues, see Zardo Trindade & Goedemé, 2016).

Figure 2. AROPE-indicator (2013) and 95% confidence intervals assuming simple random sampling.



Source: EU-SILC UDB version 2013 from 01-01-16; own calculation.
For exact values, see Appendix D.

Figure 3. Comparison of confidence intervals of Belgium and Denmark by sampling design information.



Source: Eurostat 2013, EU-SILC UDB version 2013 from 01-01-16; own calculation.
For exact values, see Appendix D.

2.3 A Stata-tool to reconstruct the sample design variables in EU-SILC

Goedemé and Zardo Trinidade have already addressed these issues in a number of papers (see e.g., Goedemé, 2010, Goedemé, 2013a, Goedemé, 2013b, Zardo Trindade & Goedemé, 2016). They also provide Stata-tools to reconstruct the missing or inappropriately coded sample design variables. The reconstruction steps take numerous country-specific modifications into account. The routine is available for the years 2005 to 2014 and can be retrieved on the homepage of Tim Goedemé.[8] Researchers who use other statistical software packages can merge the additionally provided CSV-files with the sample design variables to their data.

2.4 An SPSS-tool to reconstruct the sample design variables in EU-SILC

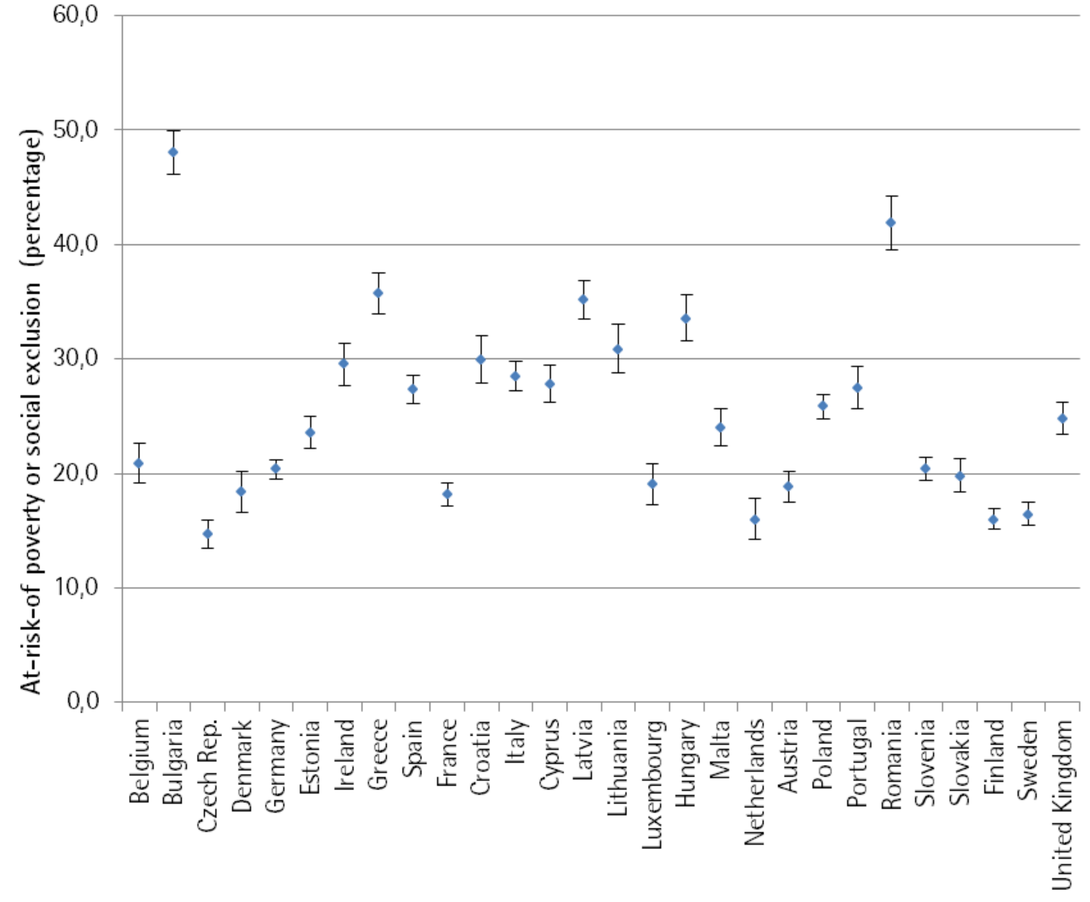
Even though researchers using statistical programs other than Stata can use the CSV-files provided by Goedemé for calculating correct standard errors, we were often asked by researchers to provide a corresponding tool for SPSS. Therefore, we have 'translated' the tool from Tim Goedemé for the EU-SILC surveys from 2010 to 2014 (cross-sectional) into SPSS. The syntax is adapted from the procedure of Goedemé and Zardo Trinidade. Thus the steps are very similar. The SPSS-syntax for the reconstruction of the sample design variables for EU-SILC 2013 is included in Appendix A. How to program SPSS for considering these variables as sampling design variables is described in Appendix B.

In Figure 4, we show the results (point estimates and the corresponding confidence intervals) for the AROPE-indicator.[9] The estimates are very close to the "true" estimates of Eurostat, using the original sample design information (Figure 1 and Table 2 in Appendix D).

[8] https://timgoedeme.com/eu-silc-standard-errors/

[9] The SPSS-syntax for calculating AROPE is included in Appendix C.

Figure 4. SPSS-Syntax: AROPE-indicator (2013) and 95% confidence intervals using sample design information (or proxies) as available in the UDB.



Source: EU-SILC UDB version 2013 from 01-01-16; own calculations using SPSS complex samples.
For exact values, see Appendix D.

2.5 Comparison of resulting confidence intervals

To illustrate the notable differences, a comparison of the confidence intervals resulting from different sampling design specifications are presented in Figure 5 to Figure 7. The figures display the confidence intervals as provided by Eurostat (2013), the ones calculated using the SPSS reconstruction syntax, as well as those calculated assuming a simple random sample design for Belgium, Bulgaria, Czech Republic, France, Greece, Italy, Poland, Romania, Spain, and the United Kingdom. These are the countries for which a specific adjustment of the reconstruction syntax was necessary (see section 2.3). The figures show that after the reconstruction of the sample design variables, the confidence intervals are close to the proper confidence intervals.

Figure 5. Comparison of the confidence intervals resulting from the different sampling design specifications by country.

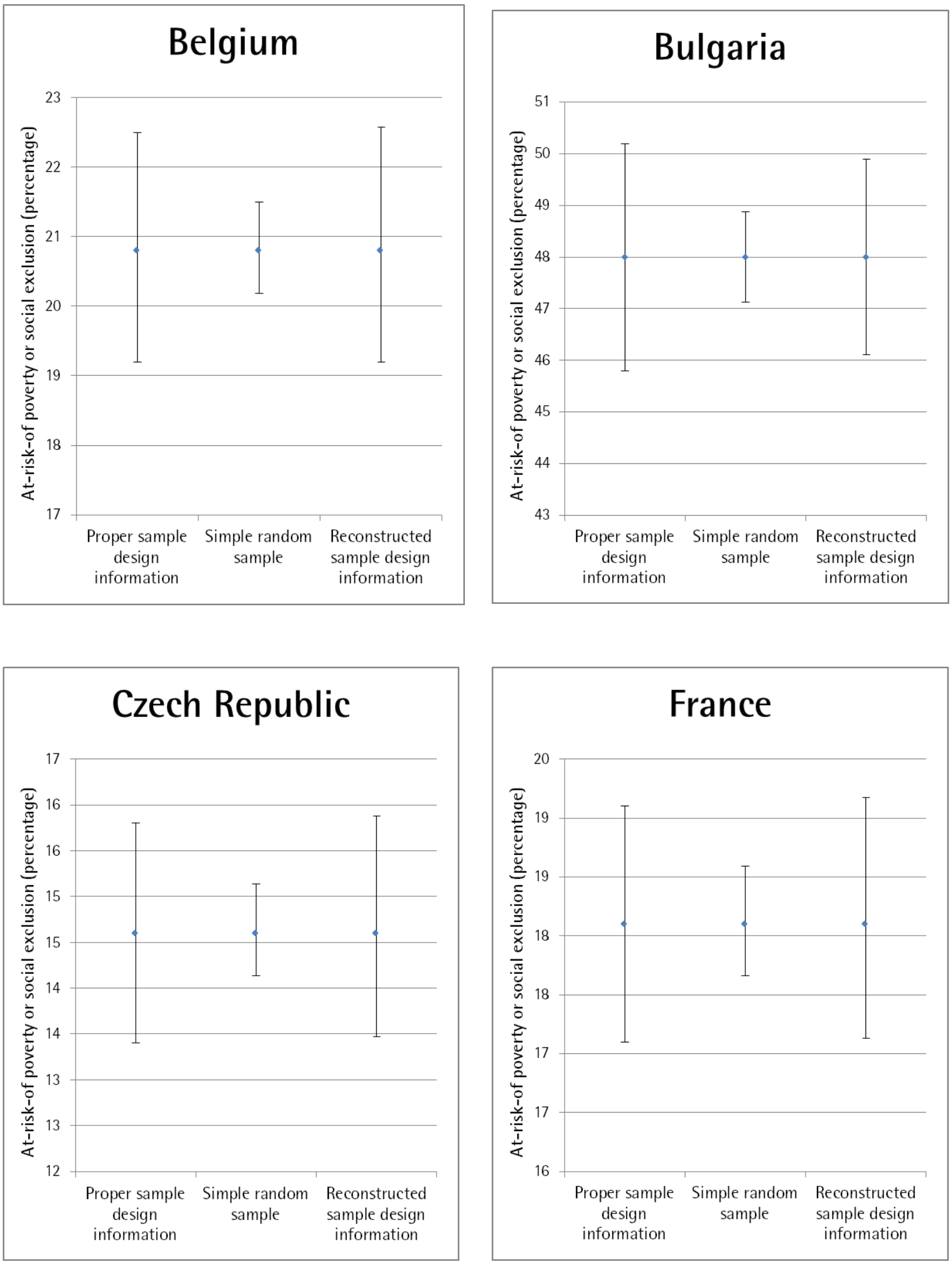
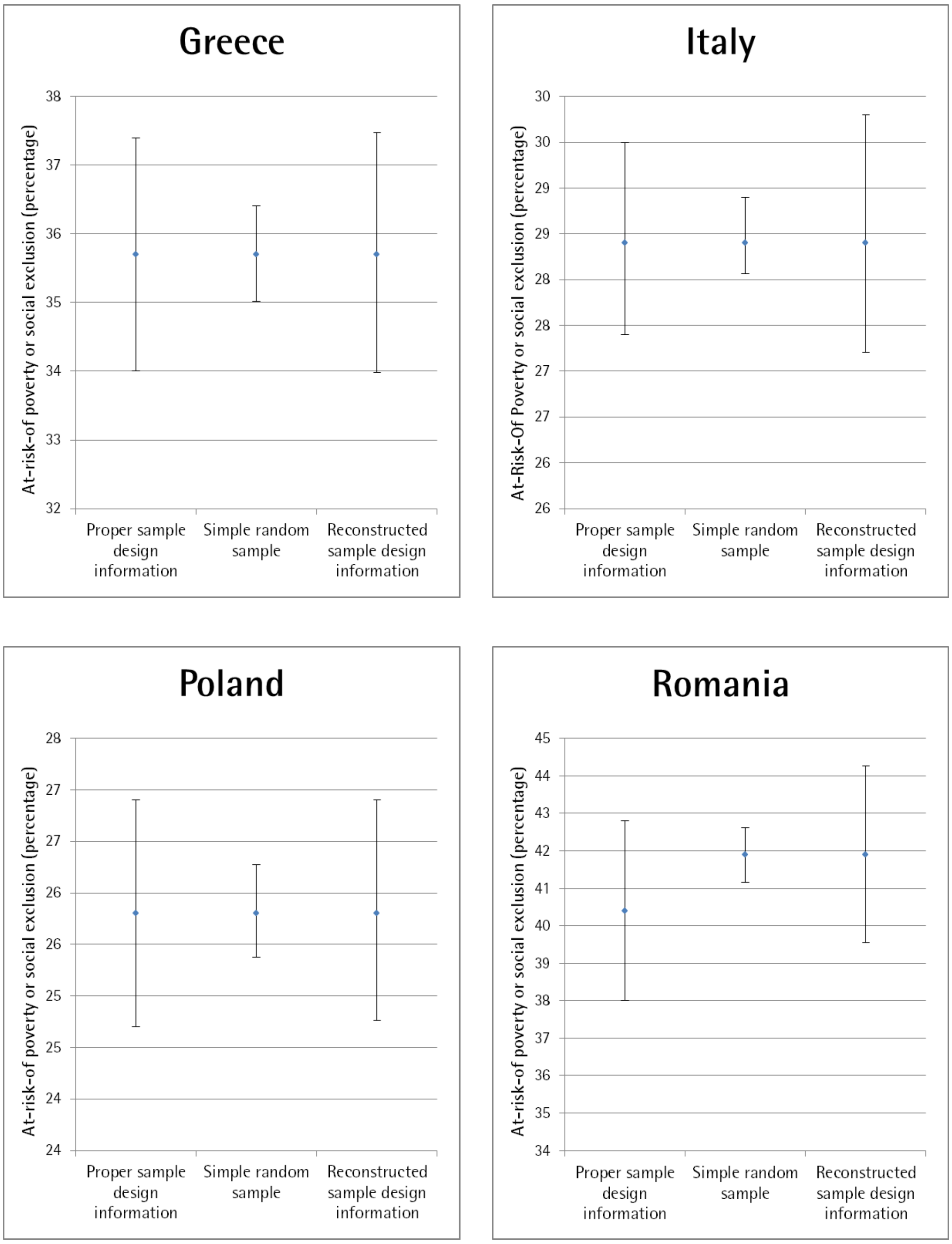
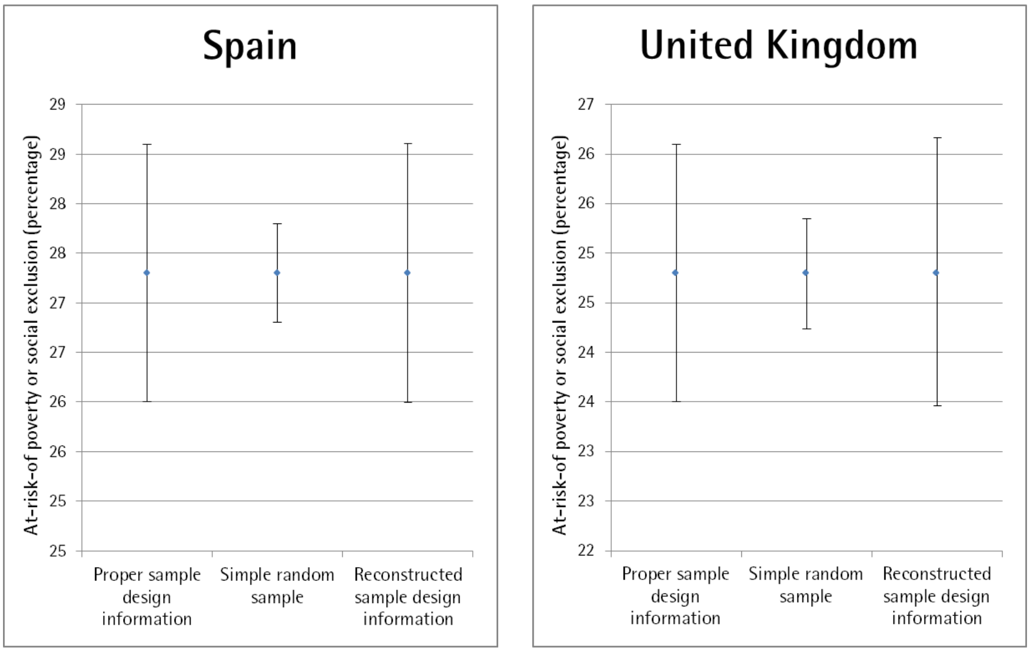


Figure 6. Comparison of the confidence intervals resulting from the different sampling design specifications by country.[10]



[10] The estimate with proper sample design information from Romania has been revised by Eurostat. According to the silc_peps01 database (Eurostat), the AROPE-estimate of Romania is 41.9.

Figure 7. Comparison of the confidence intervals resulting from the different sampling design specifications by country.



References

Atkinson, A. B., Guio, A.-C., & Marlier, E. (2017). *Monitoring social inclusion in Europ*e: Publications Office of the European Union Luxembourg.

Babbie, E. (2007). *The Practice of Social Research* (13th ed.). Wadsworth: Cengage Learning.

Córdova, A., Zéphyr, Do. (2013). Introduction to Sampling Techniques: Simple Random Samples versus Complex Samples. http://qipsr.as.uky.edu/sites/default/files/Module%201%20Sampling%20Methods.pdf. Retrieved at 2018/06/08.

Eurostat. (2013). EU comparative quality report. SILC_ESQRS_A_4D_2013_0000. Reference Metadata in ESS Standard for Quality Reports Structure (ESQRSSI). http://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports. Retrieved at: 2018/05/02.

Eurostat. (2016). Statistics explained. EU statistics on income and living conditions (EU-SILC) methodology – sampling. http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_%E2%80%93_sampling. Retrieved at: 2018/05/02.

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

Goedemé, T. (2013a). The EU-SILC sample design variables: critical review and recommendations, *CSB Working Paper* 13/02, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

Goedemé, T. (2013b). How much confidence can we have in EU-SILC? Complex sample designs and the standard error of the Europe 2020 poverty indicators. *Social Indicators Research*, 110(1), 89–110, doi:10.1007/s11205-011-9918-2.

Howes, S., & Lanjouw, J. O. (1998). Does sample design matter for poverty rate comparisons? *Review of Income and Wealth*, 44(1), 99–109.

Osier, G., Berger, Y. G. & Goedemé, T. (2013). Standard error estimation for the EU-SILC indicators of poverty and social exclusion. *Eurostat Methodologies and Working Papers Series.*

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung* (8th ed.). München: Oldenbourg Wissenschaftsverlag GmbH.

Verma, V., & Betti, G. (2006). EU Statistics on Income and Living Conditions (EU-SILC): Choosing the Survey Structure and Sample Design. *Statistics in Transition*. 7(5), 935-970.

Zardo Trindade, L. and Goedemé, T. (2016). Notes on updating the EU-SILC UDB sample design variables 2012-2014, *CSB Working Paper* 16/02, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

Appendix A. SPSS Syntax to reconstruct the sample design variables

```
* SPSS Command Syntax File
* Reconstruction of sample design variables for EU-SILC UDB 2013, version 3
*
* EU-SILC D-File (household register):
* UDB_c13D_ver 2013-3 from 01-01-16.csv
*
* For transforming the EU-SILC CSV-data (as released by Eurostat) into SPSS-data (*.sav),
* please use the corresponding SPSS Command Syntax File published at
* http://www.gesis.org/missy/eu/setups/EU-SILC
*
* (c) GESIS 07/05/2018
* GESIS - Leibniz Institute for the Social Sciences
* German Microdata Lab
* Anika Herter; Heike Wirth
* http://www.gesis.org/en/institute/
*
* Contact: heike.wirth@ gesis.org
*
***********************************************************************************************************
* AIM AND PROCEDURE
*
*    The aim of the SPSS-syntax is the reconstruction of original sample design variables psu1
*    (primary sampling unit) and strata1 (primary strata) in the D-file of the EU-SILC cross-sectional
*    data. They are necessary to compute standard errors but incomplete in the UDB.
*    The reconstruction procedure has been implemented and published for stata by Goedemé for
*     waves 2005-2011 and by Goedemé and Trinidade for waves 2012-2014. Goedemé and Trinidade
*    also documented their procedure in two articles (Goedemé 2013; Trinidade & Goedemé 2016).
*    The structure of the SPSS-Syntax at hand follows mostly the structure of the stata codes from
*    Trinidade & Goedemé.
*    The first step is some data preparation where variable DB020 is renamed as 'country' and
*     DB030 as 'hid'. Due to anonymization, the stratification variable is not at all available in the
*     UDB. The psu variable DB060 is missing for several countries. Therefore, these countries must
*     be handled individually.
*    For creating the stratification variable, it is necessary to compute a region variable for countries
*    with regional stratification (same countries as above) based on DB040. For all other countries,
*    region is set to zero. Strata0 is then computed by combining a numeric country code and the
*    previously constructed region variable. Countries with self-representing psus also have to be
*    treated individually. A 'psutest'-variable is created based on the psu-variable of the UDB (DB060).
*    For constructing the psu variable, this variable is combined with the stratification variable. For
*    countries with missing values for this variable, the household id is alternatively used. Some
*    country-specific modifications are necessary here. The variable 'checker' is constructed to identify
*    cases where households moved between selection to be a respondent and the interview itself.
*    Regrouping of psus within regions is done for these cases.
*    Finally, all not required auxiliary variables are dropped and sample design variables are reviewed.
*    SPSS Complex Samples is used to define the sampling plan (csplan) and to choose desired statistical analysis
*    (e.g. frequencies).

* For details on EU Silc standard error calculation and stata codes please visit the homepage of Tim Goedemé
* https://timGoedemé.com/eu-silc-standard-errors/
*
* Goedemé, T. (2013). '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.
* Zardo Trindade, L. and Goedemé, T. (2016) Notes on updating the EU-SILC UDB sample design variables 2012-2014,
*       CSB Working Paper 16/02, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.
*
* Please note: Since the SPSS syntax is based on the STATA syntax of Tim Goedemé
* (Download: https://timGoedemé.com/eu-silc-standard-errors/), we have also adopted the corresponding comments from him
* (marked as "Goedemé 2011")
*
***********************************************************************************************************
* Open EU Silc D-file.

set unicode no.
```

```
* - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - .

* The following command should contain the complete path and
  name of the SPSS file, usual file extension "sav";
  Change SPSS_FILENAME to your filename .

GET
  FILE='SPSS_FILENAME'.

set decimal dot.


* - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - .
* Please note:
* Until line 463 (Complex Sample Definition), there should be probably nothing to change.
*
* - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - .

****************************************************************************************************
* 1. Preparation
****************************************************************************************************
* Generate identifier.
 COMPUTE id=$CASENUM.
  FORMAT id (F8.0).
  EXECUTE.


*generate country and hid variable.
rename variables (DB020 = country).
execute.
autorecode country /into countrynr.
execute.
rename variables (DB030 = hid).

* Count number of countries.
compute numcountry = 1.
if (country eq lag(country)) numcountry = 0.
execute.
frequencies numcountry.

* show country labels.
temporary.
select if numcountry = 1.
list country .

*****************************************************************************************************
* 2. Special cases that have to be handled before rest
****************************************************************************************************


compute  psutest = DB060.

****************************************************************************************************
* 2.1 BG and PL:  "DB060 not reliable in case of Bulgaria and Poland" (Goedemé 2011: 112).
****************************************************************************************************.
do if (country='BG' or country='PL')  .
  recode psutest (0 thru 9999999 = sysmis).
end if.

****************************************************************************************************
* 2.2 FR
****************************************************************************************************.
* "France: change in sample design in 2010 complicates matters; no satisfactory method to identify
*   self-representing PSUs in old sample design" (Goedemé 2011: 114).


****************************************************************************************************
* 2.3 IT
****************************************************************************************************.
sort cases by  country DB060 DB075.
recode DB060 (SYSMIS=99999999).
```

```
compute tester=0.
execute.
do if (DB060 eq lag(DB060) and DB075 ne lag(DB075)).
        recode tester (0=1).
end if.
do if (id=1).
       recode tester (0=1).
end if.
missing values tester (0).
execute.
FREQUENCIES tester.
missing values DB060 (99999999).
missing values DB060 (99999999).
DESCRIPTIVES DB060 DB075.

sort cases by country DB060 DB075.
 AGGREGATE OUTFILE=* MODE=ADDVARIABLES
      / PRESORTED
      / BREAK  country DB060
      / npanels=sum(tester).
recode npanels (SYSMIS=0).
frequencies npanels.
sort cases by country DB060.
compute auswahl1=0.
do if (country = "IT") and (DB060 ~= lag(DB060)).
   recode auswahl1 (0=1).
end if.
filter by auswahl1.
   frequencies npanels.
filter off.

temporary.
select if ((country = "IT")).
   frequencies npanels.

do if (npanels >=2 and country = "IT").
   recode psutest (0 thru 9999999= sysmis).
end if.
temporary.
select if ((country = "IT")).
   frequencies npanels.
do if (npanels >=2 and country = "IT").
   compute tester1 = DB060.
end if.
execute.
freq tester1.
recode tester1 (sysmis = 999999).
sort cases by tester1 .
compute groupsit = 1.
do if (tester1 = lag(tester1)).
      compute groupsit= lag(groupsit).
  else if (tester1 ne lag(tester1)).
      compute groupsit = lag(groupsit) + 1.
end if.
execute.
DESCRIPTIVES groupsit.

**************************************************************************************************************
* 2.4 UK
***************************************************************************************************************.
do if country='UK'.
  compute cons=1.
end if.
execute.

missing values cons (-9).


sort cases by country DB060.
temporary.
```

```
select if (country = 'UK').
AGGREGATE OUTFILE=* MODE=ADDVARIABLES
    / PRESORTED
   / BREAK  country DB060
   / nrpsu=sum(cons).
execute.
frequencies nrpsu.

temporary.
select if (DB040 = 'UKN') .
DESCRIPTIVES nrpsu.

sort cases by country DB040.
AGGREGATE OUTFILE=* MODE=ADDVARIABLES
   / PRESORTED
   / BREAK  country  DB040
   / maximum=max(nrpsu).
execute.

do if ((country = 'UK') and (DB040 = 'UKN') and (nrpsu=maximum)).
    recode psutest (else= sysmis).
end if.
execute.

**************************************************************************************************************
* 3. Prepare Stratification variable
**************************************************************************************************************
* "AT: DB060 completely missing - therefore also stratification ignored...
*  Also for BG and PL DB060 is not sufficiently reliable -> switch to assuming a simple random
*  sample of households (probably under-estimation of the variance)" (Goedemé 2011: 182-183).

frequencies country.
string region0 (A6).
execute.
compute region0 = DB040.
*recode region0 (' ' ='0').
*execute.
Missing values region0 (' ').
execute.

* "Melilla (ES64) and Ceuta (ES63) must be grouped together as they are part of the same
*   stratum."  (Goedemé 2011: 192).
recode region0 ('ES63' 'ES64' = 'ES80').

temporary.
select if ( country= 'BE' OR country='CZ' OR country='EL'  OR country='ES'  OR country='FR'  OR country='IT'  OR country='RO').
autorecode region0 /into region0nr.
execute.
compute region1=region0nr.
recode region1 (sysmis=0).
execute.
temporary.
select if region1>0.
    descriptives region1.

sort cases by region1.


compute kons=1.
sort cases region1.
AGGREGATE OUTFILE=* MODE=ADDVARIABLES
  / presorted
  / BREAK  kons
  / regiomin = MAX(region1).
execute.
freq regiomin.

do if (npanels>=2 AND country='IT').
    COMPUTE region1a = groupsit + regiomin.
    RECODE region1a (ELSE=Copy) INTO region1.
```

```
end if .
execute.
descriptives region1.

compute strata0 = countrynr*1000 + region1.
descriptives strata0.

temporary.
select if (country='UK').
descriptives  strata0.

sort cases country strata0.
temporary.
select if (country = "UK").

AGGREGATE OUTFILE=* MODE=ADDVARIABLES
   / PRESORTED
   / BREAK kons
   / strat1=max(strata0).
execute.
compute stratum=strat1 + 2.
freq stratum.

do if country='UK' AND MISSING(psutest).
    RECODE stratum (ELSE=Copy) INTO strata0.
end if .
EXECUTE.
DESCRIPTIVES strata0.


*****************************************************************************************************************
* 4. Prepare PSU variable
*****************************************************************************************************************

COMPUTE psu0=strata0 * 100000  + hid / 100000.
EXECUTE.
DESCRIPTIVES psu0.

recode psutest (sysmis=-9999).
execute.
do if (psutest~=-9999).
     compute psu0 = strata0 * 100000 + psutest.
end if.
MISSING VALUES psutest (-9999).
DESCRIPTIVES psu0.


*****************************************************************************************************************
* 5. RE-grouping of PSUs
*****************************************************************************************************************

compute checker=1.
recode checker (1=sysmis).
execute.
sort cases by country psutest hid.

vector nocheck(1) .
execute.
rename variables nocheck1=nocheck.
recode nocheck  (sysmis = -9999).
do if (misval (psutest))OR (npanels>=2 AND country='IT').
    recode nocheck (-9999=1).
end if.
execute.
missing values nocheck (-9999).
frequencies nocheck.
missing values nocheck ().

missing values psutest ().
if ((psu0 ne LAG(psu0)) and (psutest ne lag(psutest)) OR nocheck=1)  checker=0.
```

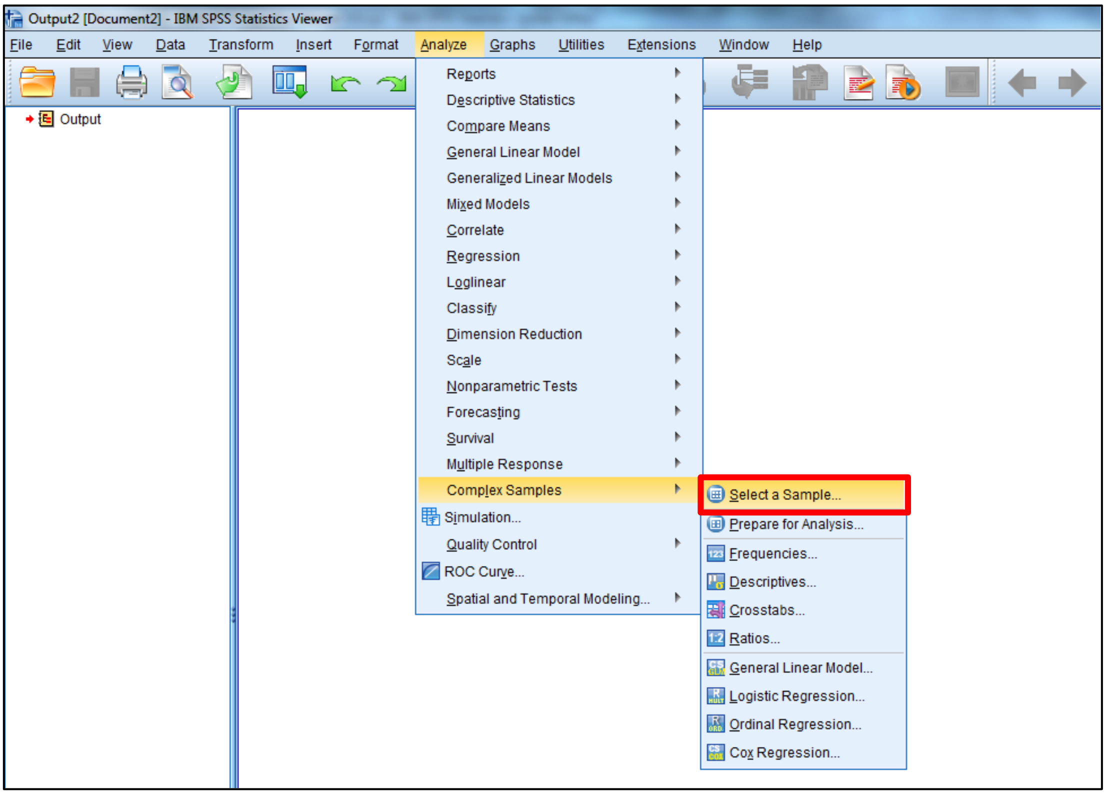
```
execute.
if ((LAG(psu0) EQ psu0) and (LAG(psutest) eq psutest) and nocheck ne 1) checker=0.
execute.
if ((LAG(psu0) ne psu0 ) and (LAG(psutest) eq psutest ) and nocheck ne 1 ) checker=1.
execute.
if ((LAG(psu0) eq psu0) and (LAG(psutest) ne psutest ) and nocheck ne 1) checker=2.
execute.
freq checker.

missing values psutest (-9999).
missing values nocheck (-9999).
frequencies nocheck.

crosstabs country by  checker.

* reset checker to 0 if PSUs must be split across strata.
do if (country ~= 'BE' and country~='CZ' and country~='ES' and country~='FR' and country~= 'IT' and country~= 'RO').
   RECODE checker (SYSMIS=0) (1=0) (2=0).
end if.
EXECUTE.
crosstabs country by checker.

sort cases by country psu0 id.

compute auswahl=0.
do if (country='BE' or country='CZ' or country='ES' or country='FR' or country= 'IT' or country= 'RO') and (psu0~=LAG(psu0)) .
   recode auswahl (0=1).
end if.
filter by auswahl.
crosstabs country by checker.
filter off.

crosstabs country by checker.


SORT CASES  BY country.
Temporary.
SPLIT FILE SEPARATE BY country.
SELECT IF ( checker = 1).
FREQUENCIES VARIABLES=psutest.
SPLIT FILE off.
Execute.


compute key=0.
sort cases country psutest strata0 id  .
AGGREGATE OUTFILE = * MODE=ADDVARIABLES OVERWRITE=YES
  / PRESORTED
  / BREAK = country psutest strata0
  / strata0freq = N(strata0).

AGGREGATE OUTFILE = * MODE=ADDVARIABLES OVERWRITE=YES
 / PRESORTED
 / BREAK = country psutest
 / aux1 = MAX( strata0freq )
 / aux2 = first(strata0)
 / aux3 = sd(strata0freq).
execute.

IF (strata0freq = aux1) key = 1.
execute.
compute strata0_mod = key*strata0.

AGGREGATE OUTFILE = * MODE=ADDVARIABLES OVERWRITE=YES
 / PRESORTED
 / BREAK = country psutest
 / checkmax = max(checker).

AGGREGATE OUTFILE = * MODE=ADDVARIABLES OVERWRITE=YES
 / PRESORTED
 / BREAK = country psutest
 / strata_01 = max(strata0_mod).


do if (checkmax=0).
  compute strata1=strata0.
end if.

do if (checkmax = 1).
  compute strata1 = strata_01.
end if.

do if (aux3=0 & checkmax=1).
  compute strata1 = aux2.
end if.
execute.

descriptives strata1 psutest psu0.

missing values psutest ().
DESCRIPTIVES psutest.

compute psu1=psu0.
execute.
do if (psutest ne -9999).
  compute psu1 =strata1*100000+psutest.
end if.
execute.
missing values psutest (-9999).

descriptives strata0 strata1 psutest psu0 psu1.

DELETE VARIABLES countrynr numcountry psutest tester npanels auswahl1 tester1 groupsit cons nrpsu maximum region0 region0nr region1 kons regiomin region1a strata0 strat1 stratum psu0 checker nocheck auswahl key strata0freq aux1 aux2 aux3 strata0_mod checkmax strata_01.
```

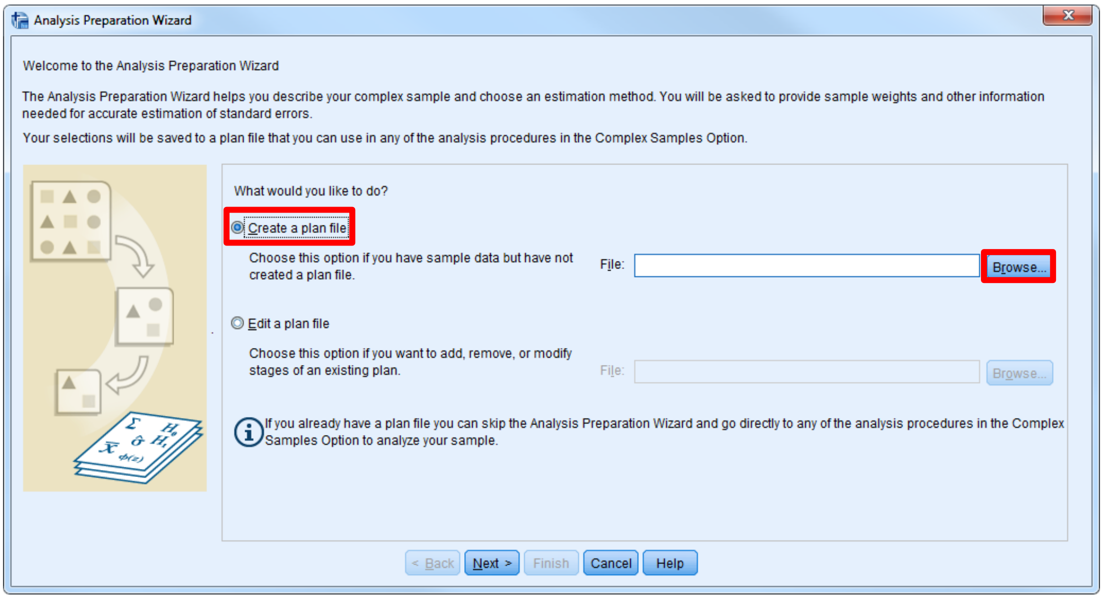
Appendix B. SPSS Syntax and instruction how to program SPSS for considering the respective variables as sampling design information

Definition of the sampling-plan: Assumptions (approximations)

- One-stage stratified random sample
- Stratification: Stratification variable (strata1)
  - 1st stage: Primary sampling units (psu1)
- Unrestricted random sample on each selection stage

SPSS Complex Samples – menu window





Welcome to the Analysis Preparation Wizard

The Analysis Preparation Wizard helps you describe your complex sample and choose an estimation method. You will be asked to provide sample weights and other information needed for accurate estimation of standard errors.

Your selections will be saved to a plan file that you can use in any of the analysis procedures in the Complex Samples Option.

What would you like to do?

● Create a plan file

Choose this option if you have sample data but have not created a plan file.

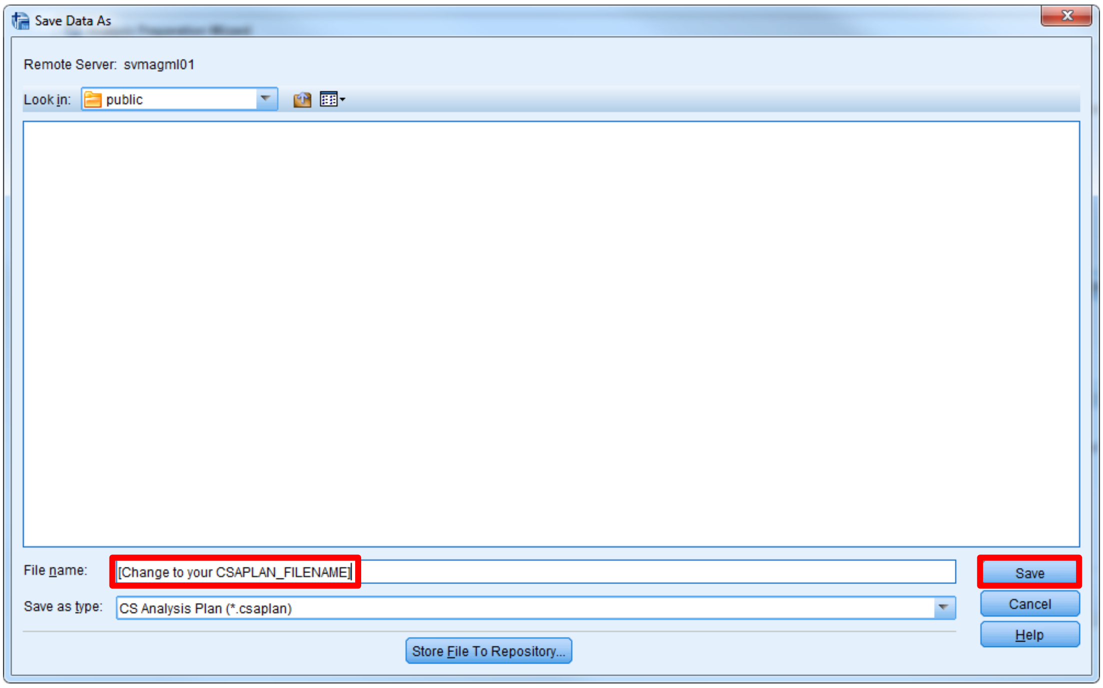
File:

○ Edit a plan file

Choose this option if you want to add, remove, or modify stages of an existing plan.

File:

If you already have a plan file you can skip the Analysis Preparation Wizard and go directly to any of the analysis procedures in the Complex Samples Option to analyze your sample.

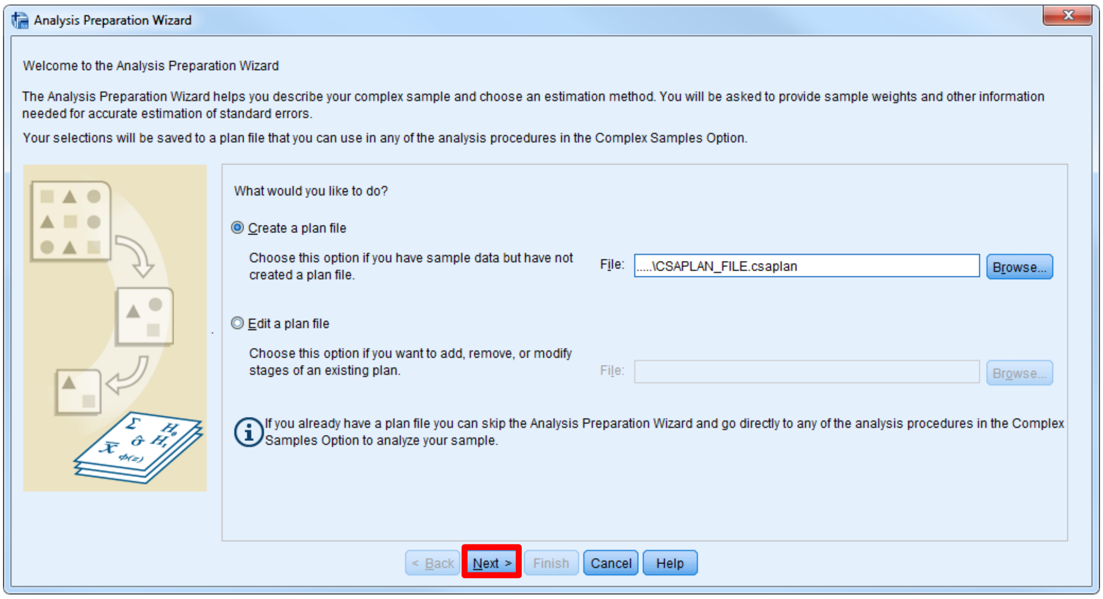


Remote Server: svmagml01

Look in: public

File name: [Change to your CSAPLAN_FILENAME]

Save as type: CS Analysis Plan (*.csaplan)



Analysis Preparation Wizard

Welcome to the Analysis Preparation Wizard

The Analysis Preparation Wizard helps you describe your complex sample and choose an estimation method. You will be asked to provide sample weights and other information needed for accurate estimation of standard errors.

Your selections will be saved to a plan file that you can use in any of the analysis procedures in the Complex Samples Option.

What would you like to do?

● Create a plan file

Choose this option if you have sample data but have not created a plan file.

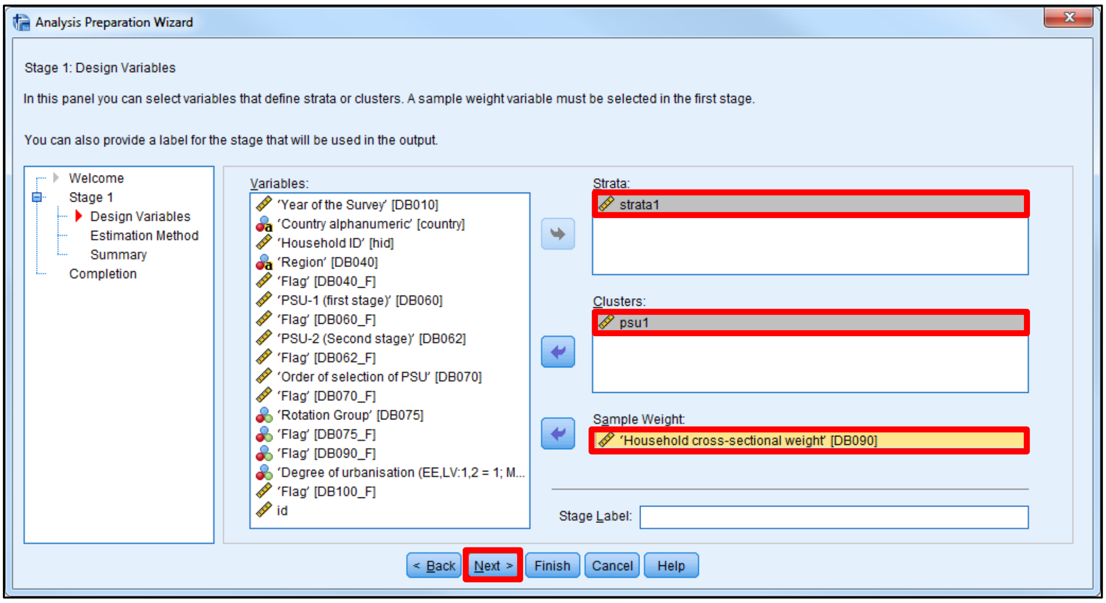
File: .....\CSAPLAN_FILE.csaplan

○ Edit a plan file

Choose this option if you want to add, remove, or modify stages of an existing plan.

File:

If you already have a plan file you can skip the Analysis Preparation Wizard and go directly to any of the analysis procedures in the Complex Samples Option to analyze your sample.



Analysis Preparation Wizard

Stage 1: Design Variables

In this panel you can select variables that define strata or clusters. A sample weight variable must be selected in the first stage.

You can also provide a label for the stage that will be used in the output.

Variables:

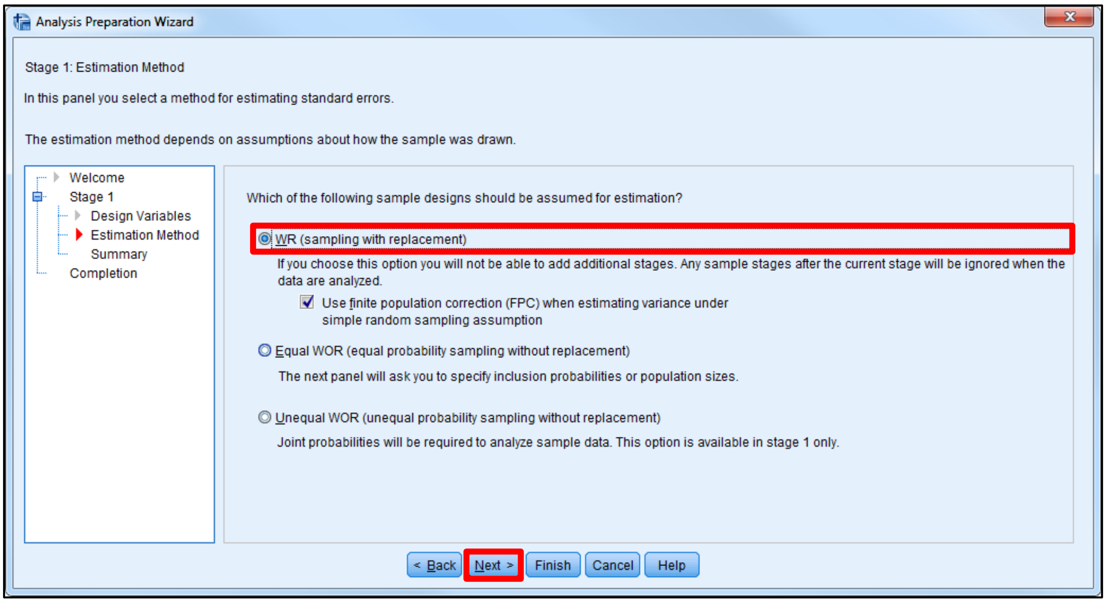
- 'Year of the Survey' [DB010]
- 'Country alphanumeric' [country]
- 'Household ID' [hid]
- 'Region' [DB040]
- 'Flag' [DB040_F]
- 'PSU-1 (first stage)' [DB060]
- 'Flag' [DB060_F]
- 'PSU-2 (Second stage)' [DB062]
- 'Flag' [DB062_F]
- 'Order of selection of PSU' [DB070]
- 'Flag' [DB070_F]
- 'Rotation Group' [DB075]
- 'Flag' [DB075_F]
- 'Flag' [DB090_F]
- 'Degree of urbanisation (EE,LV:1,2 = 1; M...
- 'Flag' [DB100_F]
- id

Strata: strata1

Clusters: psu1

Sample Weight: 'Household cross-sectional weight' [DB090]

Stage Label:



Stage 1: Estimation Method

In this panel you select a method for estimating standard errors.

The estimation method depends on assumptions about how the sample was drawn.

Which of the following sample designs should be assumed for estimation?

● WR (sampling with replacement)

If you choose this option you will not be able to add additional stages. Any sample stages after the current stage will be ignored when the data are analyzed.

☑ Use finite population correction (FPC) when estimating variance under simple random sampling assumption

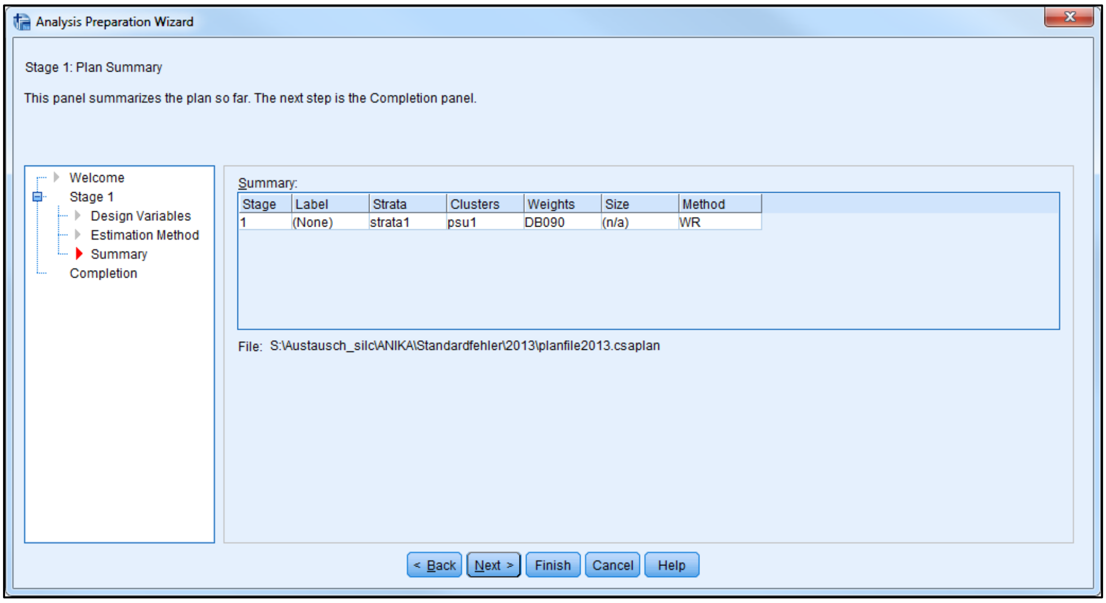
○ Equal WOR (equal probability sampling without replacement)

The next panel will ask you to specify inclusion probabilities or population sizes.

○ Unequal WOR (unequal probability sampling without replacement)

Joint probabilities will be required to analyze sample data. This option is available in stage 1 only.

Please note: Even though most countries use ‚sampling without replacement' for simplicity we assume sampling with replacement because the national inclusion probabilities are not documented.



Stage 1: Plan Summary

This panel summarizes the plan so far. The next step is the Completion panel.

Summary:

Stage	Label	Strata	Clusters	Weights	Size	Method
1	(None)	strata1	psu1	DB090	(n/a)	WR

File: S:\Austausch_silc\ANIKA\Standardfehler\2013\planfile2013.csaplan

SPSS Complex Samples – Syntax to create a plan file

```
* Analysis Preparation Wizard.
CSPLAN ANALYSIS
  /PLAN FILE='CSAPLAN_FILENAME'
  /PLANVARS ANALYSISWEIGHT=DB090
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=strata1 CLUSTER=psu1
  /ESTIMATOR TYPE=WR.
```

Appendix C. Calculating AROPE-Indicator and Complex Samples Descriptives.

Calculate AROPE-indicator based on RX070

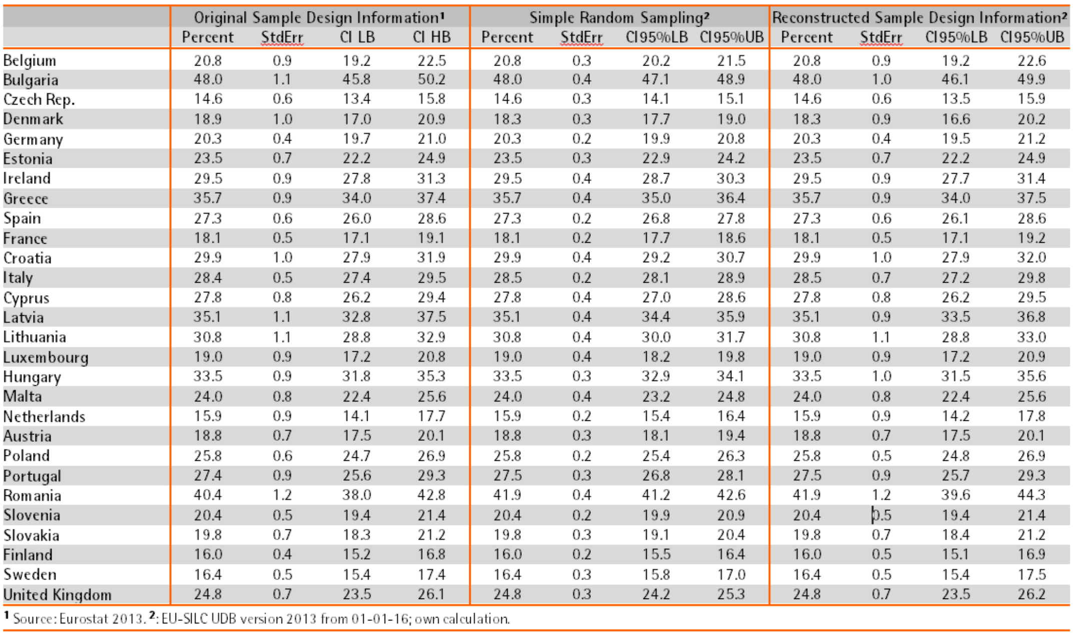
```
COMPUTE arope = 1.
do if RX070 =0.
RECODE arope  (1=0) .
end if.
```

Complex Samples Crosstabs

```
CSTABULATE
  /PLAN FILE='CSAPLAN_FILENAME'
  /TABLES VARIABLES=country BY arope
  /CELLS ROWPCT
  /STATISTICS SE CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Appendix D. Comparison of AROPE-indicator estimates considering different sample design information

Table 2. AROPE-indicator, standard error and confidence intervals by country considering different sample design information.



	Original Sample Design Information[1]				Simple Random Sampling[2]				Reconstructed Sample Design Information[2]			
	Percent	StdErr	CI LB	CI HB	Percent	StdErr	CI95%LB	CI95%UB	Percent	StdErr	CI95%LB	CI95%UB
Belgium	20.8	0.9	19.2	22.5	20.8	0.3	20.2	21.5	20.8	0.9	19.2	22.6
Bulgaria	48.0	1.1	45.8	50.2	48.0	0.4	47.1	48.9	48.0	1.0	46.1	49.9
Czech Rep.	14.6	0.6	13.4	15.8	14.6	0.3	14.1	15.1	14.6	0.6	13.5	15.9
Denmark	18.9	1.0	17.0	20.9	18.3	0.3	17.7	19.0	18.3	0.9	16.6	20.2
Germany	20.3	0.4	19.7	21.0	20.3	0.2	19.9	20.8	20.3	0.4	19.5	21.2
Estonia	23.5	0.7	22.2	24.9	23.5	0.3	22.9	24.2	23.5	0.7	22.2	24.9
Ireland	29.5	0.9	27.8	31.3	29.5	0.4	28.7	30.3	29.5	0.9	27.7	31.4
Greece	35.7	0.9	34.0	37.4	35.7	0.4	35.0	36.4	35.7	0.9	34.0	37.5
Spain	27.3	0.6	26.0	28.6	27.3	0.2	26.8	27.8	27.3	0.6	26.1	28.6
France	18.1	0.5	17.1	19.1	18.1	0.2	17.7	18.6	18.1	0.5	17.1	19.2
Croatia	29.9	1.0	27.9	31.9	29.9	0.4	29.2	30.7	29.9	1.0	27.9	32.0
Italy	28.4	0.5	27.4	29.5	28.5	0.2	28.1	28.9	28.5	0.7	27.2	29.8
Cyprus	27.8	0.8	26.2	29.4	27.8	0.4	27.0	28.6	27.8	0.8	26.2	29.5
Latvia	35.1	1.1	32.8	37.5	35.1	0.4	34.4	35.9	35.1	0.9	33.5	36.8
Lithuania	30.8	1.1	28.8	32.9	30.8	0.4	30.0	31.7	30.8	1.1	28.8	33.0
Luxembourg	19.0	0.9	17.2	20.8	19.0	0.4	18.2	19.8	19.0	0.9	17.2	20.9
Hungary	33.5	0.9	31.8	35.3	33.5	0.3	32.9	34.1	33.5	1.0	31.5	35.6
Malta	24.0	0.8	22.4	25.6	24.0	0.4	23.2	24.8	24.0	0.8	22.4	25.6
Netherlands	15.9	0.9	14.1	17.7	15.9	0.2	15.4	16.4	15.9	0.9	14.2	17.8
Austria	18.8	0.7	17.5	20.1	18.8	0.3	18.1	19.4	18.8	0.7	17.5	20.1
Poland	25.8	0.6	24.7	26.9	25.8	0.2	25.4	26.3	25.8	0.5	24.8	26.9
Portugal	27.4	0.9	25.6	29.3	27.5	0.3	26.8	28.1	27.5	0.9	25.7	29.3
Romania	40.4	1.2	38.0	42.8	41.9	0.4	41.2	42.6	41.9	1.2	39.6	44.3
Slovenia	20.4	0.5	19.4	21.4	20.4	0.2	19.9	20.9	20.4	0.5	19.4	21.4
Slovakia	19.8	0.7	18.3	21.2	19.8	0.3	19.1	20.4	19.8	0.7	18.4	21.2
Finland	16.0	0.4	15.2	16.8	16.0	0.2	15.5	16.4	16.0	0.5	15.1	16.9
Sweden	16.4	0.5	15.4	17.4	16.4	0.3	15.8	17.0	16.4	0.5	15.4	17.5
United Kingdom	24.8	0.7	23.5	26.1	24.8	0.3	24.2	25.3	24.8	0.7	23.5	26.2

[1] Source: Eurostat 2013. [2]: EU-SILC UDB version 2013 from 01-01-16; own calculation.