SSOAR

Open Access Repository

www.ssoar.info

Harmonization of Income Data in EU-SILC: Update of GESIS Paper 2015/18
Mack, Alexander; Lange, Barbara; Ponomarenko, Valentina

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der

Leibniz-Gemeinschaft

gesis   Leibniz-Institut
für Sozialwissenschaften

Harmonization of Income Data in EU-SILC

Update of GESIS Paper 2015/18

*Alexander Mack, Barbara Lange &
Valentina Ponomarenko*

Harmonization of Income Data in EU-SILC

Update of GESIS Paper 2015/18

Alexander Mack, Barbara Lange & Valentina Ponomarenko

Table of Contents

Summary

This report is intended to provide an overview of income measures employed in the EU-SILC and to inform on procedures which can be employed to harmonize these data both for comparisons over time and between countries. Each subsection provides a short overview over a topic or method and contains example code for Stata and SPSS which can be used to implement the described procedures. Alongside this report syntax files as well as a data files for both SPSS and Stata are delivered which include the code provided here as well auxiliary country level data drawn from the Eurostat database[1]. In this updated version, an extended appendix is included that explains how to generate and replicate these files in more detail.

---

[1]   These adjoining documents can be found on the MISSY pages:
      https://www.gesis.org/en/missy/materials/EU-SILC/tools/datahandling

1 Introduction[2]

The EU-SILC or European Statistics on Income and Living Conditions is a yearly data collection effort conducted by Eurostat in cooperation with European National Statistical Institutes since 2003. The primary objective of the EU-SILC is to provide comparable microdata on income, poverty, social exclusion and living conditions. Eurostat employs the EU-SILC as an important data source for indicators on income, poverty and living conditions in the EU and to evaluate progress towards EU policy objectives. Furthermore the so called EU-SILC User Database, a partially anonymized version of the microdata, is distributed to accredited researchers and has become a popular resource for comparative social research in Europe.

One of the most central tasks of the EU-SILC is to collect comparable income data. To this aim the EU-SILC followed recommendations made by the Canberra Group, an international expert group tasked with developing standards for internationally comparable household income statistics (Hoffmeyer-Zlotnik and Warner 2014: 46f.). The approach of the Canberra group proposes that income data should be collected at the household level and must consider a wide array of different income components (Canberra Group 2011).

This technical report is intended to provide an overview of income measures employed in the EU-SILC and in particular to inform on harmonization methods for comparisons over time and between countries. Each subsection provides a short overview over a topic or method and contains example code for Stata and SPSS which can be used to implement the described procedures. Further auxiliary documents are provided alongside this report.

- A Stata do-file and a SPSS sps-file which include the code detailed in this report
- A Stata and SPSS data file which include country level variables of inflation and purchasing power parity
- Stata do-files and SPSS sps-files that demonstrate how to generate and replicate the country level variables (with special files for the United Kingdom)

The remainder of the report is structured as follows: Section 2 provides information on the measurement of income data in the EU-SILC. Section 3 examines how income data can be employed for data analysis. Section 4 examines possibilities for comparing income data over time and between countries and highlights different harmonization strategies.

---

[2] An earlier version of this report was prepared as part of the Data without Boundaries project which was funded by the 7th framework programme of the European Commission.

2   Income data in the EU-SILC

A central concept within the EU-SILC is that of the income reference period. The income reference period is a fixed 12-month period, usually the previous calendar year[3]. The UK and Ireland are exceptions. In the UK the income reference period is the year of the survey while in Ireland it is the previous 12 months. All data on income collected in the EU-SILC refer to this income reference period.

Within the EU-SILC income data is collected both for households and for all working age individuals in the household. In accordance with recommendations of the Canberra Group income data are reported as components. The following sources of income are distinguished in the 2018 EU-SILC at the household level:

- imputed rent (HY030)

- income from rental of a property or land (HY040)

- family/children related allowances (HY050)

- social exclusion not elsewhere classified (HY060)

- housing allowances (HY070)

- regular inter-household cash transfers received (HY080)

- alimonies received (HY081)

- interests, dividends, profit from capital investments in unincorporated business (HY090)

- income received by people aged under 16 (HY110)

And at the individual level:
- employee cash or near cash income (PY010)

- non-cash employee income (PY020)

- company car (PY021)

- employer's social insurance contribution (PY030)

- optional employer's social insurance contributions (PY031)

- contributions to individual private pension plans (PY035)

- cash benefits or losses from self-employment (PY050)

- pension from individual private plans (PY080)

- unemployment benefits (PY090)

- old-age benefits (PY100)

---

[3]   For more information on the income variables and income reference period, consult the document "Methodological Guidelines and Description of EU-SILC Target Variables" (commonly referred to as Guidelines) page 65ff

- survivor benefits (PY110)

- sickness benefits (PY120)

- disability benefits (PY130)

- education-related allowances (PY140)

Data on these income components are reported both net and gross. Countries are required to report gross income at the individual level. However as it is up to countries whether they collect income data as gross or net information (or both) these values are sometimes imputed. On the basis of some of the income components mentioned above Eurostat calculates a number of composite income variables such as the total household gross income (HY010) or the total disposable household income (HY020). The remainder of this report will be using the total disposable household income (HY020) as a basis for illustrating different income measures.

3 Analysing income data

Household income statistics are not only difficult to collect, but presenting them in a useful and intuitive way can also be problematic as the presentation itself can decisively influence the interpretation of the given information. For different kinds of data analysis, the suitable methods and units of analysis to be used might vary. As these different units, methods and assumptions determine the conclusions that can be drawn from the results, they should be clearly stated (Canberra Group 2011: 63).

In the following examples, disposable household incomes will be the basis of all modifications and calculations. The User Database (UDB) of EU-SILC includes the variable 'HX090' which is the equivalised disposable household income. For the purpose of presenting different methods that can be used for data analysis, the starting point will be 'HY020', the total disposable household income. Multiplied with the within-household non-response factor, the resulting variable 'hydisp' will be the basis for the calculations. Further, pensions received from individual private plans (other than those covered under ESSPROS) (PY080G/PY080N) are treated as a component of property income, and should be included in the total disposable household income (HY020). From EU-SILC 2011 onwards, PY080G is automatically included in the computation of HY020, so it is preferable to add it everywhere to make data comparable. Note that all examples below use standardized identifiers, so in order for these commands to run you will have to rename your respective identifier variables into country, year and HH_ID.

Stata Code:

```
bysort country year HH_ID: egen sum080 = sum(PY080g)
gen hydisp = (HY020 + sum080) * HY025 if year < 2011
replace hydisp = HY020 * HY025 if year >= 2011
```

SPSS Syntax:

```
RECODE PY080g (SYSMIS=0) (ELSE=COPY) INTO PY080edit.
AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES
/BREAK country year HH_ID
/sum080=SUM(PY080edit).
DELETE VARIABLES PY080edit.
DO IF year < 2011.
COMPUTE hydisp = (HY020+sum080) * HY025.
ELSE IF.
COMPUTE hydisp = HY020 * HY025.
END IF.
EXECUTE.
```

3.1   Units of analysis and equivalence scales

Analysis of income data tends to focus on household income when it comes to the social or economic position of households members. The needs of a household do not grow proportionally with each additional member due to economies of scale. It is standard practice to equivalize household income by assigning weights to each household member. When applying an equivalence scale, the resulting measure is an indicator of the economic resources available to a standardized household (Canberra Group 2011: 68ff). In the most common equivalence scales, size of the household and the age of its members are taken into account in order to assign the household a value in proportion to its needs.

- The OECD equivalence scale assigns a value of 1 to the first household member, of 0.7 to each additional adult (age 14 or older) and of 0.5 to each child (age 13 or younger). This scale is sometimes referred to as the 'old OECD scale'.
- The OECD-modified scale is an adopted version of the original equivalence scale where additional adult household members (age 14 or older) are assigned a value of 0.5 (instead of 0.7) and each child (age 13 or younger) a value of 0.3 (instead of 0.5). HX090 in the EU-SILC uses this equivalence scale.

To apply these equivalence scales, the first step is to identify the number of adults (age 14 or older) and children (age 13 or younger) living in a household, as these are the criteria for the scales. Note that you must use the personal register file here as it's the only file that contains information on children.

Stata Code:
```
gen child=.

replace child=1 if RX020<14 & RX020!=.

replace child=0 if RX020>=14 & RX020!=.

gen adult=.

replace adult=1 if RX020>=14 & RX020!=.

replace adult=0 if RX020<14 & RX020!=.

bysort country year HH_ID: egen hhnbr_child = sum(child)

bysort country year HH_ID: egen hhnbr_adult = sum(adult)

gen hhnbr_pers = hhnbr_child + hhnbr_adult
```

SPSS Syntax:
```
COMPUTE adult=0.

IF (RX020>=14) adult=1.

EXECUTE.

COMPUTE child=0.

IF (RX020<14) child=1.

EXECUTE.

SORT CASES by country year HH_ID.

AGGREGATE OUTFILE=* MODE=ADDVARIABLES
```

```
/ PRESORTED

/ BREAK country year HH_ID

/ hhnbr_adult =SUM(adult)

/ hhnbr_child =SUM(child).

EXECUTE.

COMPUTE hhnbr_pers = hhnbr_child + hhnbr_adult.

EXECUTE.
```

Now, for the 'old' OECD equivalence scale, the factor 0.7 is applied for additional adults and 0.5 for children.

Stata Code:
```
gen eqs_old =.

replace eqs_old = 1+(hhnbr_adult-1) * 0.7 + hhnbr_child * 0.5 if

hhnbr_adult14>=1

replace eqs_old = 1+(hhnbr_child-1) * 0.5 if hhnbr_adult<1
```

SPSS Syntax:
```
COMPUTE eqs_old =(999).

IF (hhnbr_adult>=1) eqs_old = 1+(hhnbr_adult -1) * 0.7 +

hhnbr_child * 0.5.

IF (hhnbr_adult<1) eqs_old = 1+(hhnbr_child -1) * 0.5.

EXECUTE.
```

For the modified OECD scale, the factor 0.5 is applied for additional adults and 0.3 for children.

Stata Code:
```
gen eqs =.

replace eqs = 1+(hhnbr_adult -1) * 0.5 + hhnbr_child * 0.3 if

hhnbr_adult>=1

replace eqs = 1+(hhnbr_child -1) * 0.3 if hhnbr_adult<1
```

SPSS Syntax:
```
COMPUTE eqs = (999).

IF (hhnbr_adult>=1) eqs = 1+(hhnbr_adult -1) *0.5 +

hhnbr_child * 0.3.

IF (hhnbr_adult<1) eqs = 1+(hhnbr_child -1) * 0.3.

EXECUTE.
```

3.2   Summary measures of income level

For analysing income data, different kinds of summary measures can be suitable, depending on the research question. On the basis of macrodata or the microdata itself, income can be standard-ised using the median or the mean of the distribution. This method is commonly used given the easily accessible information that is needed. These two summary measures can lead to very differ-ent results, given that the median is less sensitive to outliers and extreme observations, unaffected by top-bottom coding and less sensitive to sampling error (Atkinson et al. 2002: 94). Both measures depend heavily on the income distribution in a country which might be especially prob-lematic for example when income distributions change specifically in certain countries (for exam-ple through economic cycles).

Both mean and median can be retrieved by using the 'summary' command and using the 'detail' option in Stata or the 'DESCRIPTIVES' option in SPSS.

Stata Code:

```
sum hydisp, detail
```

SPSS Syntax:

```
DESCRIPTIVES hydisp.

EXECUTE.
```

3.3   Measures of income dispersion

For various research questions, summary measures might not be sufficient but instead measures which asses the dispersion of income might be of relevance (Canberra Group 2011: 73f). The fol-lowing measures are the most common examples.

Frequency distribution
The frequency distribution is a very basic presentation of the distribution of income, usually visual-ised in a frequency diagram illustrating the location and spread of income within a population for each sample unit for the respective measure. As comparisons between two frequency distributions are difficult, further summary statistics are often used to complement this information.

Quantiles
This approach is based on the ranking of units of analysis according to ascending income. Here, the population of interest is divided into equally sized groups, called quantiles, which capture the share of total income accruing to a given proportion of households or persons. Depending on one's research interests there are different ways to employ quantiles. When dividing the popula-tion into five equally sized groups, the quantiles are called quintiles. For ten groups they are re-ferred to as deciles or percentiles for 100 groups. For example, when examining the first quintile of a distribution, the result is a measure of the income distribution among the bottom 20 percent of the population along the income variable of interest. This can then be compared with other measures like the mean or the median of the whole population to get a clearer impression of the circumstances of the respective group. Calculating quantiles is one way of making data compara-ble over time and also across countries, as it only considers the relative positioning within a coun-try and at a certain point of time.

The following example generates a variable where the five quintiles of the income distribution are saved for all countries. In Stata the country variable has to be encoded (country_num). Depending on how many countries are contained in your sample you have to redefine the range of i in the below forvalues command. In SPSS the procedure is more straightforward.

Stata Code:

```
encode country, gen(country_num)

gen quintile=.

forvalues i =1/25 {

xtile quintile`i'= hydisp if country_num==`i', nq(5)

replace quintile=quintile`i' if country_num==`i'

drop quintile`i'

}
```

SPSS Syntax:

```
RANK VARIABLES = hydisp BY country /NTILES(5) INTO quintile.
```

4   Cross-time and cross-country comparisons

In order to be able to compare income data over time and across countries these data must be standardized in some shape or form. The example of quantiles above achieves comparability by converting a measure of income into a measure of relative position within the income distribution. Other procedures are more geared towards making income units comparable either between countries by adjusting for purchasing power or over time by factoring out inflation.

4.1   Adjusting for price changes over time: Price level indices

For comparisons over time, income data has to be adjusted using an appropriate price index that is consistent with the income definition and refers to the same population group. For a narrow definition of income, for example net disposable income, the price index should only capture those consumption items that can in fact be purchased out of net disposable income; for broader definitions of income, the range of included items in the price level index should be correspondingly broader as well (Canberra Group 2011: 84).

One widely used concept is that of harmonised indices of consumer prices (HICP) which is available from Eurostat[4] and is calculated to reflect household monetary consumption expenditure in the EU. Along with this technical report, an inflation file is delivered that can be merged to the EU-SILC data. It contains inflation weights for all countries for all years using the 2003[5] data as a basis (as this is the income reference period of the first survey year of the EU-SILC). For Switzerland and Serbia the reference years differ. The HICP is only available for Switzerland as of 2005 and for Serbia as of 2006. Note that for all countries the indices of the preceding year are used for the inflation weight, as this is the income reference year in the EU-SILC[6]. The inflation weight is constructed as the relative growth of inflation compared to the reference year.

The dataset and corresponding .sps-files and do-files can be downloaded from the missy website[7]. For more details please consult the appendix. Using the following command, this aggregate dataset will be merged to your EU-SILC microdata. Please make sure that the needed identifiers 'year' and 'country' are named correspondingly (also, make sure that Greece is coded 'EL' in your data). The code below assumes that you have saved inflppps file in your C: directory.

Stata Code:

```
merge m:1 year country using "C:\inflppps.dta"
drop if _merge==2
```

---

[4]   prc_hicp_aind in the Eurostat database

[5]   In an earlier version it was wrongly stated that the inflation weight is based on 2004, as it is the first survey year. The first survey of 2004 corresponds to the income reference period of 2003.

[6]   The UK uses the year of the survey year as the income reference period. The infl_weight variable accounts for this and includes values for the actual survey year for UK. In the case of Ireland the income reference period are the 12 months prior to the survey thus technically it is not possible to apply proper inflation weighting with yearly data. The infl_weight variable includes the values for the previous year for Ireland, however this operationalization lacks precision

[7]   https://www.gesis.org/en/missy/materials/EU-SILC/tools/datahandling [Accessed 11 February 2020]

SPSS Syntax:

```
SORT CASES BY country year.
MATCH FILES FILE= *
/TABLE =' C:\inflppps.sav'
/BY country year.
EXECUTE.
```

Years and countries that are not in your data are deleted by dropping those cases that are not matched from the original dataset. The weighting is then applied by dividing the respective income variable by the inflation weight. However note that the resulting values are converted to the income level of 2004.

Stata Code:

```
gen inflwhydisp=hydisp/infl_weight
```

SPSS Syntax:

```
COMPUTE inflwhydisp=hydisp/infl_weight.
EXECUTE.
```

4.2 Comparison between geographical areas

For comparisons across geographical areas income data should be adjusted to account for differences in price levels. Purchasing power parities (PPPs) are a measure of the relative prices of goods and services that are likely to be purchased for consumption purposes, for example by households given their household income, and take account of differences in national price levels and allow for comparisons in real terms (Canberra Group 2011: 86). PPPs are calculated comparing price levels for a basket of comparable goods and services that are selected to be representative of consumption patterns in the various countries. PPPs can thus be used to convert income data into an artificial common currency.

The calculation of PPPs is at its core based on expenditure data, but as they reflect the purchasing power of money, they can be applied to income measures as well. The choice of the appropriate PPP depends on the measure of income that is of interest. When comparing household income data, PPPs based on households' consumption expenditure should be applied (Canberra Group 2014: 173ff). The purchasing power parities (calculated in Euro) for household consumption expenditure[8] are also included in the delivered country level file as the variable 'ppps'. With these PPPs, income data in the EU-SILC can be converted to a common synthetic currency which is comparable between countries for a given year.

---

[8] prc_ppp_ind in the Eurostat Database

As basis for the common synthetic currency, EU28 countries are included. An additional file with PPPs on grounds of EU27 without the United Kingdom can be computed as well. More information is presented in the appendix.

In order to merge the ppps variable to your EU-SILC data follow the instructions detailed in section 4.1. Similar to the infl_weight variable values in the ppps variable correspond to the income reference period and refer to the previous year for all countries save for the UK. In order to harmonize income variables they must be divided by the ppps[9] variable.

Stata Code:

```
gen pppshydisp=hydisp/ppps
```

SPSS Syntax:

```
COMPUTE pppshydisp=hydisp/ppps.
```

---

[9]   Values of the ppps in the current version of the file differ slightly from previous versions. This is because data revisions of PPPs time series took place in 2016. For more information consider ESA 2010 Regulation 549/2013.

5 References

Atkinson, A. B. (2002). Social indicators: The EU and social inclusion. Oxford University Press.

Canberra Group (2011). Handbook of Household Income Statistics. Second Edition 2011. Geneva: United Nations.

Eurostat (2013). Description of Target Variables: Cross-sectional and Longitudinal 2012 operation.

Hoffmeyer-Zlotnik, J. H., & Warner, U. (2014). Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research. Springer.

Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013 on the European system of national and regional accounts in the European Union Text with EEA relevance [Online]. [Accessed 11 February 2020]

Available from https://eur-lex.europa.eu/

6 Appendix

6.1   How to the generate *infl_weight*

The variable *infl_weight* is generated using the data sheet "Harmonised indices of consumer prices (HICP)" provided by Eurostat. The raw data in form of comma separated values (*.csv) needs to be downloaded from the EUROSTAT website[10]. The indicator needs to be obtained in a certain form and with specific settings. The settings are summarized in the box below. As of now, HICP are provided on basis of the values of 2015. In order to replicate and extend *infl_weight*, annual data has to be chosen.

For the dataset, please make the following settings: Choose the unit of measure as annual average index. Choose from the classification of individual consumption by purpose (COICOP): All-items HICP (CP00). You may adjust the year and country selection. Please be aware that the values for *infl_weight* can only be produced if Eurostat provides HICP for the country at a certain year. Data for Switzerland is available from 2005 and for Serbia from 2006 onwards. Choose in the download settings: cell formatting in the form "1 234.56".

Box 1: Download and formatting settings: Harmonised indices of consumer prices (HICP)

Title: HICP (2015=100)-annual data (average index and rate of change)

Abbreviation: prc_hicp_aind

Unit of measure: Annual average index (INX_A_AVG)

Classification of individual consumption by purpose (COICOP): All-items HICP (CP00)

Time: 2003-2017

Geo: AT, BE, BG, CH, CZ, CY, DE, DK, EE, EL, ES, FI, FR, HR, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, UK

After saving the csv dataset, please load the sps-file or do-file "infl_weight.*" and specify your paths. Only the dataset prc_hicp_aind _1_Data.csv is then necessary. If you don't want to include the United Kingdom, use infl_weight_without_UK.*.

---

[10]   http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_hicp_aind [Accessed 11 February 2020]

6.2   How to the generate *ppps*

The variable *ppps* is generated using the data sheet "Purchasing power parities (PPPs)" provided by Eurostat. The raw data in form of comma separated values (*.csv) needs to be downloaded from the EUROSTAT website[11]. The indicator needs to be obtained in a certain form and with specific settings. The settings are summarized in the box below. The purchasing power parities by Eurostat have been revised in 2016. Therefore, it is advised to download the current data file from the website. In order to replicate and extend the variable *ppps*, the indicator needs to address the price level indices and real expenditures for ESA 2010 aggregate.

For the dataset, please make the following settings: Choose as national accounts indicator, Purchasing power parities (EU28=1). Choose from the analytical categories for purchasing power parities calculation: Actual individual consumption (A01). You may adjust the year and country selection. Choose in the download settings cell formatting in the form "1 234.56".

Box 2: Download and formatting settings: Purchasing power parities (PPP)

> Title: Purchasing power parities (PPP), price level indices and real expenditures for ESA 2010 aggregates
>
> Abbreviation: prc_ppp_ind
>
> National accounts indicator (NA_ITEM): Purchasing power parities (EU28=1)
>
> Analytical categories for purchasing power parities calculation (PPP_CAT): Actual individual consumption (A01)
>
> Time: 2003-2017
>
> Geo: AT, BE, BG, CH, CZ, CY, DE, DK, EE, EL, ES, FI, FR, HR, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, UK

If you want to generate ppps without accounting for the United Kingdom, choose for national accounts indicator Purchasing power parities (EU27_2020=1). Proceed with ppps_without_UK.*.

---

[11]   http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_ppp_ind [Accessed 11 February 2020]

Another data sheet needs to be downloaded to convert the non-Euro currencies to Euro. This is done with Eurostat's Euro/ECU exchange rates[12]. Make the following settings (Box 3) and download the data. Choose in the download settings cell formatting in the form "1 234.56".

Box 3: Download and formatting settings: Euro/ECU exchange rates

---

Title: Eurostat: Euro/ECU exchange rates - annual data

Abbreviation: ert_bil_eur_a

Unit of measure (UNIT): national currency

Statistical Information (STATINFO): Average

Time: 2003-2018

Currencies: Bulgarian lev, Czech koruna, Swiss franc, Danish krone, Croatian kuna, Hungarian forint, Icelandic krona, Norwegian krone, Polish zloty, Romanian leu, Serbian dinar, Swedish krona, Pound sterling

---

After saving the csv dataset, please load the do-file or sps-file "ppps.*" and specify your paths. Only the dataset prc_ppp_ind _1_Data.csv and ert_bil_eur_a_1_Data.csv are then necessary.

---

[12]   http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ert_bil_eur_a [Accessed 11 February 2020]