



AUTOMATIC DATA VALIDATION FOR SILC

Margherita Zuppardo

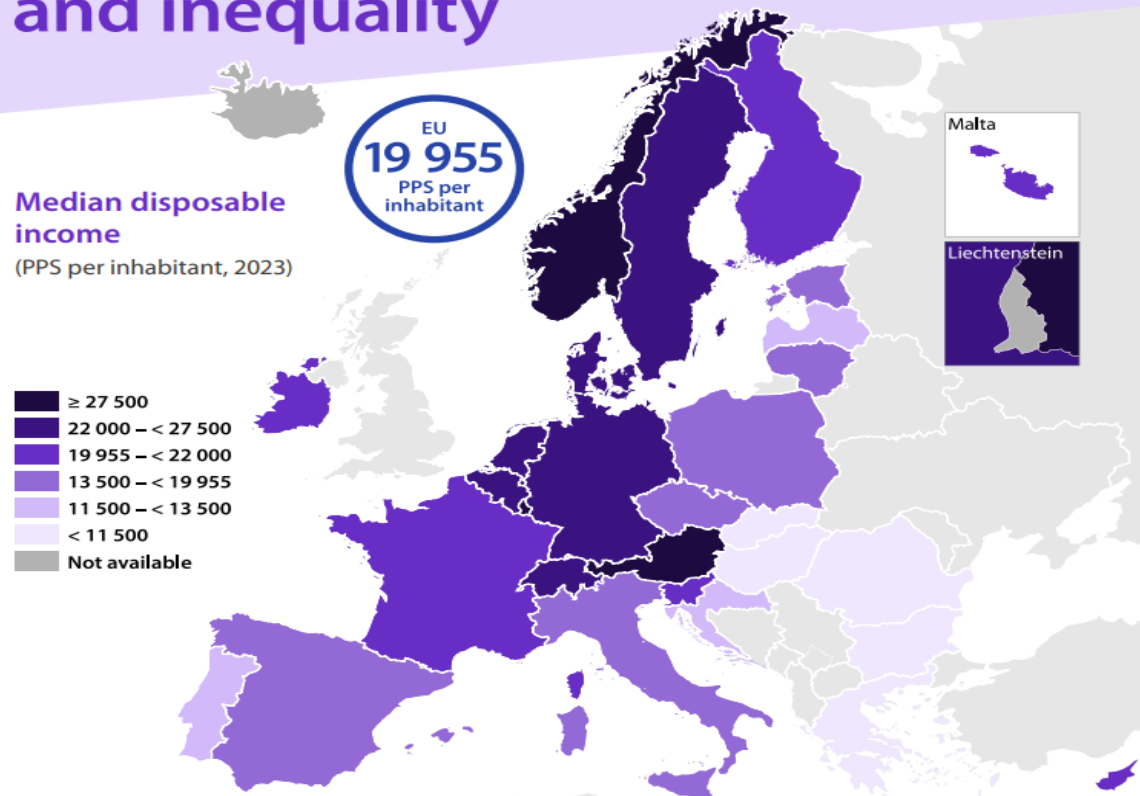
margherita.zuppardo@hagstofa.is



Statistics Iceland

What is SILC?

Income distribution and inequality



The EU statistics on income and living conditions (EU-SILC) aim to collect timely and comparable cross-sectional and longitudinal data on **income, poverty, social exclusion, and living conditions**.

EU-SILC is a household and individual data collection which output is harmonised as it is regulated by legislations. Around 90% of the data collection is made up of annual variables. The rest are either

[1,2] ec.europa.eu

The SILC Team in Iceland

Data collection



Data processing and analysis



EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social Statistics
Unit F-4: Quality of life

The “doc65”

DocSILC065 (2023 operation)

METHODOLOGICAL GUIDELINES AND DESCRIPTION OF EU-SILC TARGET VARIABLES

2023 operation (Version 5)

SILC Variables

TABLE OF CONTENTS

PART I: EU-SILC GUIDELINES

1. Introduction.....
2. Standardised and core variables.....

3. Technical specifications.....

4. General information.....

5. Income and expenditure.....

6. Flags.....

7. Data collection.....

8. Weights.....

9. Timeliness.....

PART II: DEFINITION OF HOUSEHOLD

1. DB010: Year of birth.....

2. DB020: Country of birth.....

3. DB030: Country of residence.....

4. DB040: Region of residence.....

5. DB050: Strata.....

6. DB060: Priority.....

7. DB062: Sex.....

8. DB070: Order.....

9. DB075: Role.....

10. DB076: Interview.....

11. DB080: Household.....

12. DB090: Household.....

13. DB095: Household.....

HS060: Capacity to face unexpected expenses.....

HS090: Do you have a computer.....

HS110: Do you have a car.....

HS120: Ability to make a trip.....

HS150: Financial burden.....

HD080: Replacing work.....

HI010: Change in the household.....

HI020: Reason for income change.....

HI030: Reason for decision.....

HI040: Expectation of income.....

HY010: Total household income.....

HY020: Total disposable income.....

HY022: Total disposable income and survivor's benefits.....

HY023: Total disposable income and survivor's benefits.....

HY040G/HY040N: Income.....

HY050G/HY050N: Family income.....

HY051G: Family/child income.....

HY052G: Family/child income.....

HY053G: Family/child income.....

HY054G: Family/child income.....

HY060G/HY060N: Social exclusion.....

HY061G: Social exclusion.....

HY062G: Social exclusion.....

HY063G: Social exclusion.....

HY064G: Social exclusion.....

HY070G/HY070N: Housing allowance.....

HY071G: Housing allowance.....

HY072G: Housing allowance.....

HS011: Arrears on hire purchase instalment.....

HS021: Arrears on hire purchase instalment.....

HS022: Reduced payment on hire purchase instalment.....

HS031: Arrears on hire purchase instalment.....

HS040: Capacity to afford paying for one year.....

HS050: Capacity to afford a meal with meat, chicken, fish or vegetables once a day.....

RB083: Passing of birthday at the time of the interview.....251

RB090: Sex.....252

RB100: Sample person or co-resident.....

RB110: Membership status.....

RB120: Location where the person lives.....

RB200: Residential status.....

RB211: Main activity status (self-defined).....

RB220: Father ID.....

RB230: Mother ID.....

RB240: Spouse/partner ID.....

RB245: Respondent ID.....

RB250: Data source.....

RB280: Country of birth.....

RB285: Duration of stay.....

RB290: Country of residence.....

RL010: Educational level.....

RL020: Educational level.....

RL030: Childcare.....

RL040: Childcare.....

RL050: Childcare.....

RL060: Childcare.....

RL070: Childcare.....

RG_Z#; Grid.....

PERSONAL DATA.....

PB010: Year of birth.....

PB020: Country of birth.....

PB030: Person ID.....

PB040: Person ID.....

PB050: Person ID.....

PB060: Person ID.....

PB070: Person ID.....

PB080: Person ID.....

PB081: Age in completed years.....

PB082: Age in completed years.....

PD050: Get-together with friends/family [relatives] for a drink/meal.....month.....

PD060: Regularly participate in a leisure activity.....

PD070: Spend a small amount of money each month.....

PD080: Internet connection for personal use.....

PW010: Overall life satisfaction.....

PW191: Trust in others.....

PL016: Existence of previous partner.....

PL032: Self-defined current partner.....

PL040A: Status in employment.....

PL040B: Status in employment.....

PL051A: Occupation in main job.....

PL051B: Occupation (last job).....

PL060: Number of hours usually worked.....

PL073: Number of months since last job.....

PL074: Number of months since last job.....

PL075: Number of months since last job.....

PL076: Number of months since last job.....

PL080: Number of months since last job.....

PL085: Number of months since last job.....

PL086: Number of months since last job.....

PL087: Number of months since last job.....

PL088: Number of months since last job.....

PL089: Number of months since last job.....

PL090: Number of months since last job.....

PL100: Total number of hours worked.....

PL111A: Economic activity.....

PL111B: Economic activity.....

PL141: Permanency of main job.....

PL145: Full or part-time main job.....

PH010: Self-perceived health.....

PH020: Suffer from any chronic [long-standing] illness.....

PH030: Limitation in activities because of health problems.....

PH040: Unmet need for medical examination or treatment.....

PH050: Main reason for unmet need for medical examination or treatment.....

PH060: Unmet need for dental examination or treatment.....

PH070: Main reason for unmet need for dental examination or treatment.....

PD020: Replace worn-out clothes by some new (not second-hand) ones.....

PL150: Super.....

PL200: Number.....

PL211A: Main.....

PL211B: Main.....

PL211C: Main.....

PL211D: Main.....

PL211E: Main.....

PL211F: Main.....

PL211G: Main.....

PL211H: Main.....

PL211I: Main.....

PL211J: Main.....

PL211K: Main.....

PL211L: Main.....

PL271: Duration.....

PY010G/PY01.....

PY020G/PY02.....

PY021G/PY02.....

PY030G: Employment.....

PY035G/PY03.....

PY050G/PY05.....

PY080G/PY08.....

PY090G/PY09.....

PY091G: Unemployment.....

PY092G: Unemployment.....

PY093G: Unemployment.....

PY094G: Unemployment benefits [non-contributory and non-means-tested].....

PY100G/PY100N: Old-age benefits.....

PY101G: Old-age benefits (Contributory and means-tested).....

PY102G: Old-age benefits (Contributory and non means-tested).....

PY103G: Old-age benefits (Non-contributory and means-tested).....

HD140: One meal with meat, chicken or fish.....day.....

HD150: Books at home suitable for their age.....

HD160: Outdoor leisure equipment.....

HD170: Indoor games.....

HD180: Regular leisure activity.....

HD190: Celebration on special occasions.....

HD200: Invite friends round to play or eat from.....

HD210: Participate in school trips and school.....

HD220: Suitable place to study or do homework.....

HD240: Go on holiday away from home at least.....

AD-HOC SUBJECT MODULE 2021.....

HK010: Number of children who are household.....outside the household.....

HK020: Number of children who are not household.....inside the household.....

PK010: Household member has children who.....

RK010: Parent ID and sequential number of.....

RK020: Age of the child who is not household.....

PK020: Main reason for not spending more time.....members.....

PK030: Main reason for not spending more time.....members.....

RK030: Usual time the parent needs to get to.....member.....

RK040: Frequency of contact (via phone, social.....with the child who is not a household member.....

RK050: Child having a bedroom in which to.....siblings).....

RK060: Frequency of spending actively time.....walking, talking etc.).....

RK070: Number of nights per month the child.....

RK080: Legal child custody situation.....

RCH010: General health.....

RCH020: Limitation in activities because of health problems (child).....

HCH010: Unmet need for medical examination or treatment (child).....

HCH020: Main reason for unmet need for medical examination or.....

HCH030: Unmet need for dental examination or treatment (child).....

HCH040: Main reason for unmet need for dental examination or treatment.....

HD100: Some new (not second-hand) clothes.....

HD110: Two pairs of properly fitting shoes (including a pair of all).....

HD120: Fruits and vegetables once a day.....

Data processing in R

```
year1='21' #year I am working on
```

```
#make the D-file
```

```
source('D_file_light.r')
```

```
#make the R-file
```

```
source('R_file.r')
```

```
#make the P-file
```

```
source('P_file_tax.r')
```

```
##make the H-file
```

```
source('H_file_no_tax.r')
```

```
source(paste('ad_hoc',year1,'.r',sep=''))
```

```
source(paste('rolling_modules_20',year1,'.r',sep=''))
```

```
source('make_reconciled2.r')
```

```
source('weights_integrated_bergro22.r')
```

```
if(year1 != year1_start) error= erro
```

```
source('Reorder_files3.r')
```

Data processing in R

```
year1='21' #year I am working on
```

```
#make the D-file
```

```
source('D_file_light.r')
```

```
#make the R-file
```

```
source('R_file.r')
```

```
#make the P-file
```

```
source('P_file_tax.r')
```

```
##make the H-file
```

```
source('H_file_no_tax.r')
```

```
source(paste('ad_hoc',year1,'.r',sep=''))
```

```
source(paste('rolling_modules_20',year1,'.r',sep=''))
```

```
source('make_reconciled2.r')
```

```
source('weights_integrated_bergros22.r')
```

```
if(year1 != year1_start) error= erro
```

```
source('Reorder_files3.r')
```

```
#####check codes
```

```
#General checks on the variables
```

```
source('check_values.r')
```

```
#Aggragate consistency among years
```

```
source('check_values_weights.r')
```

```
#Total populations stemming from the weights
```

```
source('check_weights.r')
```

```
#Logical consistency checks (between variables and between years)
```

```
source('logical_checks.r')
```

Data validation

```
#####check codes
```

```
#General checks on the variables
```

```
source('check_values.r')
```

```
## missing columns
```

```
## too many missing values (above 20%)
```

```
## accepted values
```

```
## correct flags
```

```
#Aggragate consistency among years
```

```
source('check_values_weights.r')
```

```
## mean, median and sd of weighted variables consistent with previous year
```

```
#Total populations stemming from the weights
```

```
source('check_weights.r')
```

```
#Logical consistency checks (between variables and between years)
```

```
source('logical_checks.r')
```


Data validation

```
#####check codes
```

```
#General checks on the variables
```

```
source('check_values.r')
```

```
## missing columns
```

```
## too many missing values (above 20%)
```

```
## accepted values
```

```
## correct flags
```

```
#Aggragate consistency among years
```

```
source('check_values_weights.r')
```

```
## mean, median and sd of weighted variables consistent with previous year
```

```
#Total populations stemming from the weights
```

```
source('check_weights.r')
```

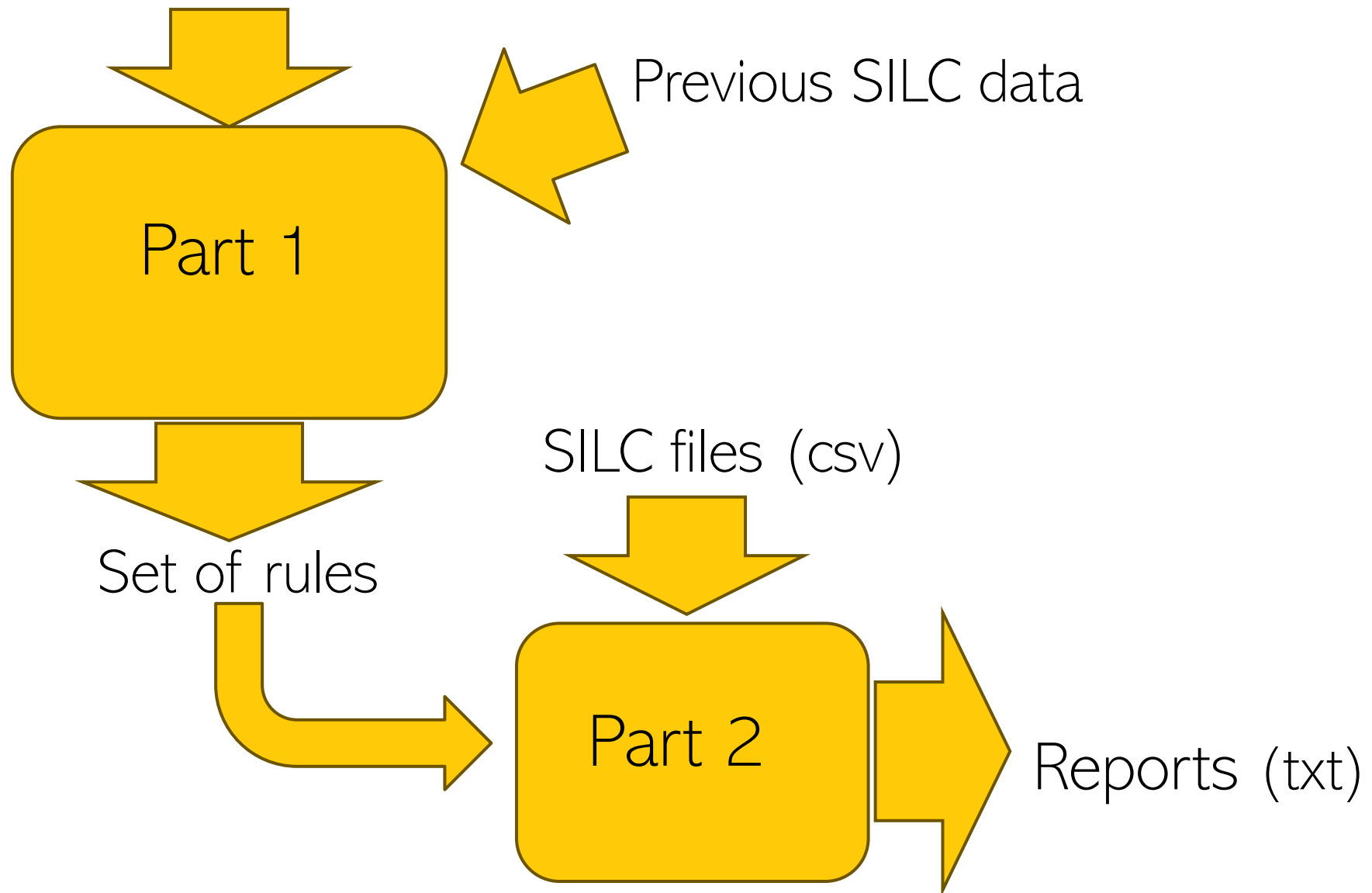
```
#Logical consistency checks (between variables and between years)
```

```
source('logical_checks.r')
```

.

Eurostat documents (pdf, excel)

Our code



Why we need automatic rule extraction

“Rules” change between years

RB090: Sex

Domain/Area	Basic data/Demograph	
Transmission type	Early and regular	
Reference period	Constant	
Unit	All current household members (of any age) and former household members	
Mode of collection	Household respondent	
Values	1	Male
	2	Female
Flags	1	Filled
	-1	Missing

```
(R_file1 |> filter(RB090_F %!in% c(1,-1)) |> nrow
```

2021

RB090: SEX

Topic and detailed topic: Person and household characteristics / Demography
Variable type: Core Variable/First wave/ Annual
Unit: All current household members (of any age)
Reference period: Current
Mode of collection: Household respondent or registers
In use (period): Yes, since the first year of EU-SILC data collection
Series' differences: No changes

VALUES AND FORMAT

- 1 Male
- 2 Female

FLAGS

- 1 Main source is survey or interview
- 2 Main source is administrative data
- 3 Imputed
- 4 It is not possible to establish a main source
- 1 Missing

```
(R_file1 |> filter(RB090_F %!in% c(1,2,3,4,-1)) |> nrow()) >0
```

library(pdftools)

Household Data (H-file)

030: REASON FOR DECREASE IN INCOME

Topic and detailed topic: Income, consumption and elements of wealth, including debts/ Total annual income at household and respondent level

Variable type: Annual

Unit: Household

Reference period: Current

Mode of collection: Household respondent

In use (period): Part of 2019 ad-hoc module, collected annually from 2021

Series' differences: Yes, in 2021 moved to annual

VALUES AND FORMAT

- 1 Reduced working time, wage or salary (same job), including self-employment (involuntary)
- 2 Parenthood/ parental leave /child care/ to take care of a person with illness or disability
- 3 Changed job
- 4 Lost job/unemployment/ bankruptcy of (own) enterprise
- 5 Became unable to work because of illness or disability
- 6 Divorce / partnership ended / other change in household composition
- 7 Retirement
- 8 Cut in social benefits
- 9 Other

FLAGS

- 1 Filled
- 1 Missing
- 2 Not applicable
- 7 Not applicable

DESCRIPTION

Step 1: pdf text

HY010: TOTAL HOUSEHOLD GROSS INCOME

Topic and detailed topic: Income, consumption and elements of wealth, including debts/ Total annual income at household and respondent level

Variable type: Annual

Unit: Household

Reference period: Income reference period

Mode of collection: Derived

In use (period): Yes, since first year of EU-SILC data collection

Series' differences: No changes

VALUES AND FORMAT

-999999.99 - 999999.99 Income (national currency) without inflation factor

FLAGS

Type of variable	Flag name	Type and content	Type of information	Values
Income variable	_F	Two-digit flag: first digit	Most common source or method	1 Collected via survey/interview
				2 Collected from administrative data
				3 Deductive/logical imputation (also including top and bottom-coding)
				4 Gross/net conversion
				5 Model-based imputation
				6 Donor imputation
	Two-digit flag: second digit	Type of collected value		7 Not possible to establish the most common source or method
				1 Net of tax on income at source and social contributions
				2 Net of tax on income at source
				3 Net of social contributions
				4 Mix of different nets
				5 Gross
				6 Income component(s) not taxed and gross
				7 (the value was not collected)
				999999.99-999999.99
				dividing by 0 appears

```
# Read the text from the selected PDF file
```

```
pdf_text <- pdf_text(pdf_path)
```

```
pdf_data <- tibble(page_number = seq_along(pdf_text), text = pdf_text)
```

[3] <https://cran.r-project.org/web/packages/pdftools>

Sections

```
#### now add the 'topic' column which says what the line is describing
mutate(
  # Identify lines that contain specific keywords
  is_values = str_detect(lines, paste0("^\\s*", keywords$Values_string, "\\s*$")),
  is_flags = str_detect(lines, paste0("^\\s*", keywords$Flags_string, "\\s*$")),
  is_description = str_detect(lines, paste0("^\\s*", keywords$Description_string, "\\s*$")),

  # Generate section ids based on occurrences of these keywords
  section_id = cumsum(is_values | is_flags | is_description)
) %>% ungroup() %>%
# Assign topic based on the section type detected
mutate(
  topic = case_when(
    is_values ~ "values",
    is_flags ~ "flags",
    is_description ~ "description", # Optionally h
    TRUE ~ NA_character_
  )
) %>%
```

Household Data (H-file)

HY010: TOTAL HOUSEHOLD GROSS INCOME

Topic and detailed topic: Income, consumption and elements of wealth, including debts / Total annual income at household and respondent level

Variable type: Annual

Unit: Household

Reference period: Income reference period

Mode of collection: Derived

In use (period): Yes, since first year of EU-SILC data collection

Series' differences: No changes

VALUES AND FORMAT

999999.99 - 999999.99 Income (national currency) without inflation factor

FLAGS

HI030: REASON FOR DECREASE IN INCOME

Household Data (H-file)

Topic and detailed topic: Income, consumption and elements of wealth, including debts/ Total annual income at household and respondent level

Variable type: Annual

Unit: Household

Reference period: Current

Mode of collection: Household respondent

In use (period): Part of 2019 ad-hoc module, collected annually from 2021

Series' differences: Yes, in 2021 moved to annual

VALUES AND FORMAT

- 1 Reduced working time / wage or salary (same job), including self-employment (involuntary)
- 2 Parenthood/ parental leave /child care/ to take care of a person with illness or disability
- 3 Changed job
- 4 Lost job/unemployment/ bankruptcy of (own) enterprise
- 5 Became unable to work because of illness or disability
- 6 Divorce / partnership ended / other change in household composition
- 7 Retirement
- 8 Other social benefits
- 9 Other

FLAGS

- 1 Filled
- 1 Missing
- 2 Not applicable (HI010 not equal to 3)
- 7 Not applicable (HB010 not equal to 2021)

DESCRIPTION

Type of variable	Flag name	Type and content	Type of information	Values
Income variable	_F	Two-digit flag: first digit	Most common source or method	1 Collected via survey/interview
				2 Collected from administrative data
				3 Deductive/logical imputation (also including top- and bottom-coding)
				4 Gross/net conversion
				5 Model-based imputation
				6 Donor imputation
				7 Not possible to establish the most common source or method
				1 Net of tax on income at source and social contributions
				2 Net of tax on income at source
				3 Net of social contributions
	_IF	Two-digit flag: second digit	Type of collected value	4 Mix of different nets
				5 Gross
				6 Income component(s) not taxed
				7 Mix of net and gross
				8 Unknown
				9 Not applicable (the value was not collected)
				-999999.99-999999.99
				Imputation factor = collected value / recorded value *100
				- If problem of dividing by 0 appears

```
range_pattern <-
```

```
"(-?\d+(?:\.\d+)?)\s*(?:[---]|to)\s*(-?\d+(?:\.\d+)?)"
```

```
(-?\d+(?:\.\d+)?)
```

A negative integer, or a positive integer, or a real number

```
\s*(?:[---]|to)\s*
```

A hyphen (-), en-dash (—), or em-dash (—), the word „to“

VALUES AND FORMAT

0–99 Number of hours of education during a typical week

VALUES AND FORMAT

1 - 9999999.99 Benefits
- 9999999.99 - -1 Losses

VALUES AND FORMAT

1–31 Day

VALUES AND FORMAT

BE	Belgium
BG	Bulgaria
CZ	Czechia
DK	Denmark
DE	Germany

Format: Four-digit number, no decimals

VALUES AND FORMAT

Household from previous wave	
1	At the same address
2	Entire household moved
Household no longer in-scope	
3	Entire household moved
4	Household moved o

VALUES AND FORMAT

NUTS 2 digits

```
library(stringr)
```

Extracting the rule[4]

```
var = 'HY010'  
text <- pdf_data |> filter(variable_name == var & topic == 'values' & lines != keywords$values_string)  
pull(lines) |> paste0()
```

HY010: TOTAL HOUSEHOLD GROSS INCOME

```
# Detect ranges like "0-9", or "-9999.999 - 99999.99", or "1.5 to 3.5"  
ranges <- stringr::str_match_all(text, range_pattern)  
print(ranges)
```

```
> print(ranges)  
[[1]]  
      [,1]      [,2]      [,3]  
[1,] "-999999.99 - 999999.99" "-999999.99" "999999.99"
```

```
# Transforms numbers to logical conditions  
if (length(ranges[[1]]) > 0) {  
  condition <- paste('value >= ', ranges[[1]][, 2], " & value <= ", ranges[[1]][, 3])  
}  
print(condition)
```

Result

```
> print(condition)  
[1] "value >= -999999.99 & value <= 999999.99"
```

Unit: Household
Reference period: Income reference period
Mode of collection: Derived
In use (period): Yes, since first year of EU-SILC data collection
Series' differences: No changes

VALUES AND FORMAT

-999999.99 - 999999.99 Income (national currency) without inflation factor

FLAGS

Values

1	Collected via survey/interview
2	Collected from administrative data
3	Deductive/logical imputation (a and bottom-coding)
4	Gross/net conversion
5	Model-based imputation
6	Donor imputation
7	Not possible to establish the source or method
1	Net of tax on income at source contributions
2	Net of tax on income at source
3	Net of social contributions
4	Mix of different nets
5	Gross
6	Income component(s) not taxed
7	Mix of net and gross

flag: first digit source or method

flag: second digit type of collected value

Other validation

```
#####check codes
```

```
#General checks on the variables
```

```
source('check_values.r')
```

```
## missing columns
```

```
## too many missing values (above 20%)
```

```
## accepted values
```

```
## correct flags
```

```
#Aggragate consistency among years
```

```
source('check_values_weights.r')
```

```
## mean, median and sd of weighted variables consistent with previous
```

```
#Total populations stemming from the weights
```

```
source('check_weights.r')
```

```
#Logical consistency checks (between variables and between years)
```

```
source('logical_checks.r')
```









<https://cran.r-project.org/web/packages/rvest>

<https://cran.r-project.org/web/packages/pxweb>

<https://cran.r-project.org/web/packages/readxl>

Final reports

Name

-  report_D-file2021
-  report_H-file2021
-  report_H-file2021_weighted
-  report_logical_checks2021
-  report_P-file2021
-  report_P-file2021_weighted
-  report_R-file2021
-  weights_errors_2021

Final reports

There are 4 variables with errors or warnings in R-file

There are 2 variables with too many missing values in R-file

There are 1 variables with values that are not allowed in R-file

There are 1 variables with wrong flags in R-file

```
***** RK020*****  
                AGE OF CHILD WHO IS NOT HOUSEHOLD MEMBER
```

Warning! Variable: RK020. 1 values outside the desired range. Example: 98

```
***** RK070*****  
                NUMBER OF NIGHTS PER MONTH THE CHILD SPENDS IN THE HOUS
```

Warning! Variable RK070 has 22 % missing values.

```
***** RK080*****  
                LEGAL CHILD CUSTODY SITUATION
```

Warning! Variable RK080 has 22 % missing values.

```
***** RB245*****  
                RESPONDENT STATUS
```

Years valid: >= 2006 . Error: Flag values NA not allowed. Expected value: 1.



report_logical_checks2021 - Notepad

Final reports

File Edit View

----- as.numeric(PY010G)<= 0 & (PL211A == 1 | PL211A == 2) -----

----- Person is an employee (PL211A = 1 or 2) but has no income from employment (PY010G == 0) -----

29 row(s) affected

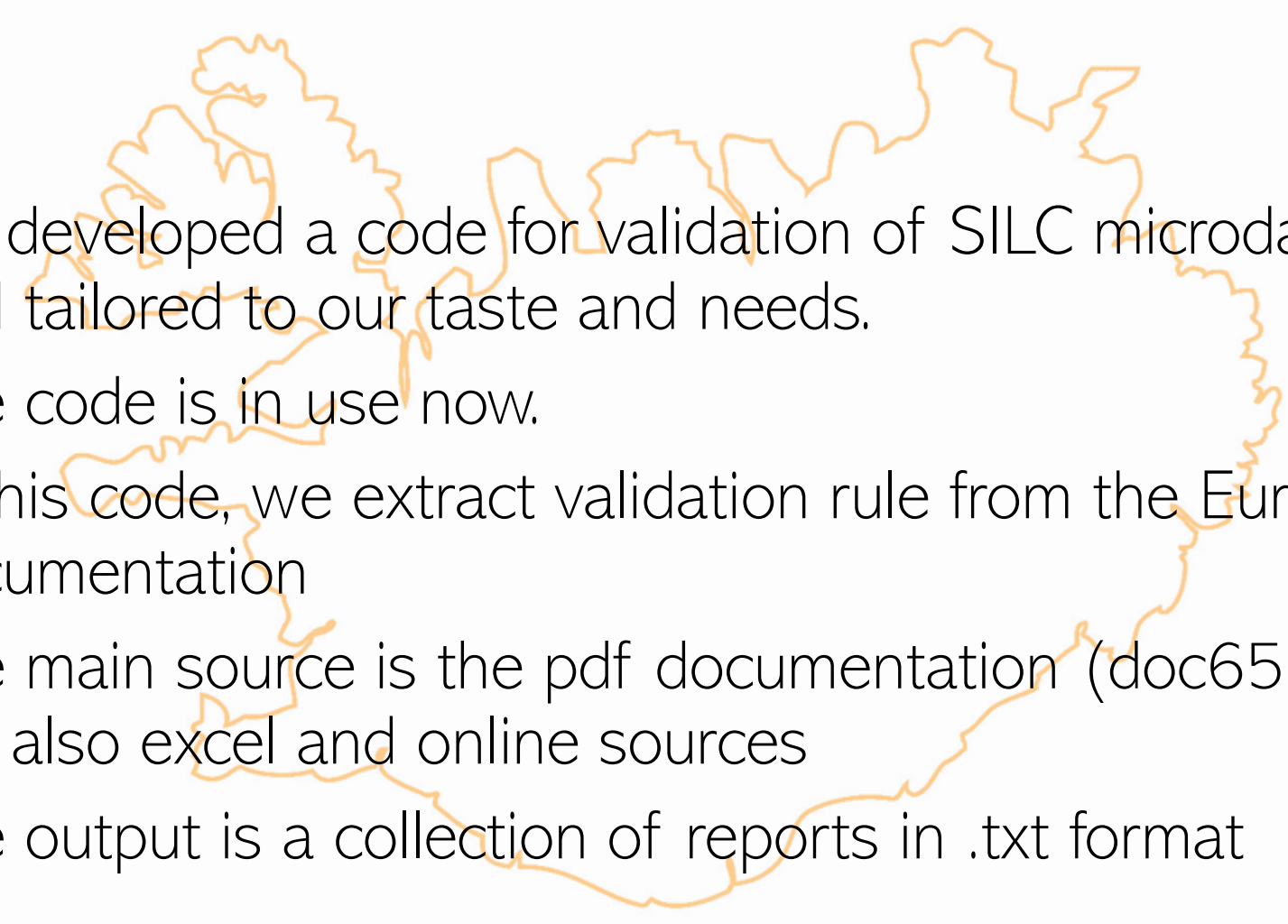
"Temp_id2"	"PY010G"	"PY010G_F"	"PL211A"	"PL211A_F"
"03471700012020"	"0"	"0"	"2"	"1"
"05341700012020"	"0"	"0"	"1"	"1"
"08451700012020"	"0"	"0"	"2"	"1"
"09331700012020"	"0"	"0"	"1"	"1"
"02811800012020"	"0"	"0"	"1"	"1"
"04311700012020"	"0"	"0"	"1"	"1"
"12321800012020"	"0"	"0"	"1"	"1"
"02181800012020"	"0"	"0"	"1"	"1"
"03041900012020"	"0"	"0"	"1"	"1"
"05041900012020"	"0"	"0"	"2"	"1"

----- as.integer(PL060) +as.integer(PL100) >60 -----

----- Nr of hours usually worked is more than 60 hours per week -----

64 row(s) affected

Conclusions

- 
- We developed a code for validation of SILC microdata and tailored to our taste and needs.
 - The code is in use now.
 - In this code, we extract validation rule from the Eurostat documentation
 - The main source is the pdf documentation (doc65), but also excel and online sources
 - The output is a collection of reports in .txt format

Pros:

- It works (most of the time)
- R-based pipeline simplifies SILC validation.
- The validation works on different years, with little maintenance.
- It takes minutes to run
- Automation reduces human error
- All packages (and R) are free

Conclusions

Cons:

- Other programming languages (Python) may offer more tools
 - It took a while to develop the code initially.
- It depends on Eurostat documentation uniformity, requires robust design.
 - It still needs some manual coding (logical checks)
 - It is dependent on external packages
 - Not as complete as Eurostat SAS code



THANK YOU!

- [1] ec.europa.eu/eurostat/web/products-key-figures/w/ks-01-24-001
- [2] ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions
- [3] <https://cran.r-project.org/web/packages/pdftools>
- [4] <https://cran.r-project.org/web/packages/stringr/>

Code coming soon

<https://github.com/MargheritaZ/SilcValidation>



Statistics Iceland

margherita.zuppardo@hagstofa.is

```
range_pattern <- "(\\d+(?:\\.\\d+)?|\\-\\d+(?:\\.\\d+)?)\\s*(?:[---]|to)\\s*(\\d+(?:\\.\\d+)?|\\-\\d+(?:\\.\\d+)?)"
discrete_pattern <- "(?<=^|\\n)\\s*(\\d+)\\.?\\s+(?!-)(?!\\.)*"
open_range_pattern <- "(\\d+)\\+"
```