|  |  |  |
| --- | --- | --- |
| logo_ec_17_colors_300dpi | EUROPEAN COMMISSION  EUROSTAT  Directorate F: Social statistics  **Unit F-4: Income and living conditions; Quality of life** |  |

|  |
| --- |
| **EU-SILC Validation Software**  **USER GUIDE – EG and SAS 9.4 Users**  **Ver 4.0** |

**Table of contents**

[**Introduction** 3](#_Toc179792124)

[**1.** **Preliminary information** 4](#_Toc179792125)

[**1.1** **Methodological references** 4](#_Toc179792126)

[**1.2** **SILC operation years** 4](#_Toc179792127)

[**1.3** **SILC Data Files** 4](#_Toc179792128)

[**1.4** **Useful operational notations** 5](#_Toc179792129)

[**2.** **Running the EU-SILC VAL SW** 5](#_Toc179792130)

[**2.1** **SAS EG** 5](#_Toc179792131)

[**2.1** **SAS SA** 6](#_Toc179792132)

[**3.** **The validation reports** 7](#_Toc179792133)

[**3.1** **Overview reports** 8](#_Toc179792134)

[**3.2** **Detailed reports** 8](#_Toc179792135)

[**3.3 Analysis reports** 9](#_Toc179792136)

[**3.4 Reports’ analysis and interpretation** 9](#_Toc179792137)

[**ANNEX: The VAL SW structure** 13](#_Toc179792138)

# **Introduction**

Following the entry into force of IESS [Regulation (EU) 2019/1700 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R1700&from=EN), in 2021 Eurostat has re-engineered the internal IT tool to validate the data of the European Union Survey on Income and living conditions (EU-SILC) and to calculate the related indicators, publicly available on the Eurostat’s website.

Pursuing annex V of the IESS Regulation, Member States shall submit for the Income and Living Conditions domain pre-checked microdata without direct identifiers. Therefore, as also agreed during the EU-SILC Working Group held on 13-14 December 2022, Eurostat shares the VAL SW with NSIs in CIRCABC, requiring SAS not older than version 9.4.

The developed Validation software (that in this document will be labelled as VAL SW) consists in a SAS Project containing several SAS programs that run the validation checks on SILC data and generate reports in different formats. The tool has been designed and developed for SAS Enterprise Guide (EG) users as well as for SAS Stand Alone (SA) users.

Please note that pre-validation carried out with other statistical packages/programming languages (R, Python, Julia, SQL, etc.) is also allowed.

On 22.10.2024, Eurostat released version 4.0 of the VAL SW:

* in the CIRCABC directory available [here](https://circabc.europa.eu/ui/group/853b48e6-a00f-4d22-87db-c40bafd0161d/library/f73b2850-af6d-4101-93f3-70347723e6a6?p=1&n=10&sort=modified_DESC).
* in the github repository <https://github.com/eurostat/EU-SILC>

Eurostat will inform NSIs in case any update occurs in the software and in the accompanying documentation.

The SW is labelled “DATACHECKXML.egp”.

This User Guide aims at providing EU-SILC data validators (internal Eurostat’s EU-SILC production team members and colleagues from the NSIs) with practical, operational information on the EU-SILC VAL SW.

The document, suitable for both SAS EG and SAS SA Users, presents the following outline:

1. Information on methodological references, operation years, SILC data files and notation to be followed to properly execute the validation
2. Instruction on how to run the VAL SW for both EG and SA Users
3. Instruction on how to analyse and interpret the validation reports
4. Annex 1 reports an overview of the SAS programs that perform the validation process

Eurostat invites colleagues from NSIs to contact the EU-SILC production team ([ESTAT-ILC@ec.europa.eu](mailto:ESTAT-ILC@ec.europa.eu)) for any information needed on the presented VAL SW.

# **Preliminary information**

# **Methodological references**

For the User’s information, the SW is based on a number of validation rules checking the compliance of SILC variables, values and flags reported by data providers in the SILC files with the reference variables/flags and relative modalities contained in a XML file. Such file is annually configured according to the EU-SILC Methodological guidelines (DOC065).

# **SILC operation years**

As SILC provides cross-sectional and longitudinal information[[1]](#footnote-1), the SW is configured to refer to the proper validation rules for each operation year.

In other terms, if a country in December 2024 provides a 4-year panel SILC data, its data file will report information for the last 4 years (2021, 2022, 2023 and 2024 – or back to 2019 in case of 6 year panel).

The VAL SW will validate the files according to the methodological rules applied in each different operation year. Following the entry into force of the IESS Regulation in 2021, compared to previous edition, the EU-SILC methodological guidelines underwent several changes: revision of variables’ identifiers and/or modalities, changes in the flags, introduction of new variables, implementation of recurring modules, shift of previous optional/ad-hoc variables to nucleus, etc.

A “disruption year” can be identified in 2021, when the Regulation requirements were implemented at methodological level in the EU-SILC exercise. Therefore, the validation process/outcomes related to operation years until 2020 (included) are marked with the “ANTE” prefix, while the validation material regarding operation years as from 2021 are marked with the “POST” prefix.

# **SILC Data Files**

NSIs send EU-SILC data in four separate files:

* Household register (D)
* Personal register (R)
* Household data (H)
* Personal data (P)

The household register file (D) contains every household selected, including those where the address could not be contacted and households that could not be interviewed.

In the other files, records associated with a household will exist only if the household has been contacted and has a completed household interview in the household data file (H) and at least one member has complete data in the personal data file (P). This member must be the selected respondent if this mode of selection is used.

The personal register file (R) contains a record for every person currently living in the household or temporarily absent. As regards the longitudinal component, it must also contain a record for every person registered in the previous year’s R-file.

Finally, the personal data file (P) contains a record for every eligible person for whom the information could be completed from interview and/or registers.

# **Useful operational notations**

As mentioned in the installation guide, **&eusilc/main/&cc/csv** is the inbound directory where the SILC data files to be processed should be stored. They have to comply with the following naming rules:

* + SILC\_**R&F**\_A\_**&cc**\_**&yyyy**\_0000\_V**&FVER**.csv

Where “&FVER” is a 4 digit code. For example, the SILC 2023 D file version 1 for Austria is labelled as follows: SILC\_D\_A\_AT\_2024\_0000\_V0001.csv.

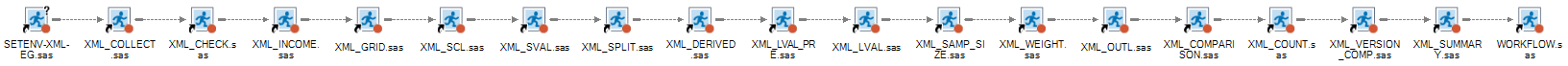
Please bear in mind that all SILC data files (D, R, H, P) to be validated must have the same version number.

The validation report generated by the SW will be automatically stored in the output directory **&eusilc/main/&cc/out** .

# **Running the EU-SILC VAL SW**

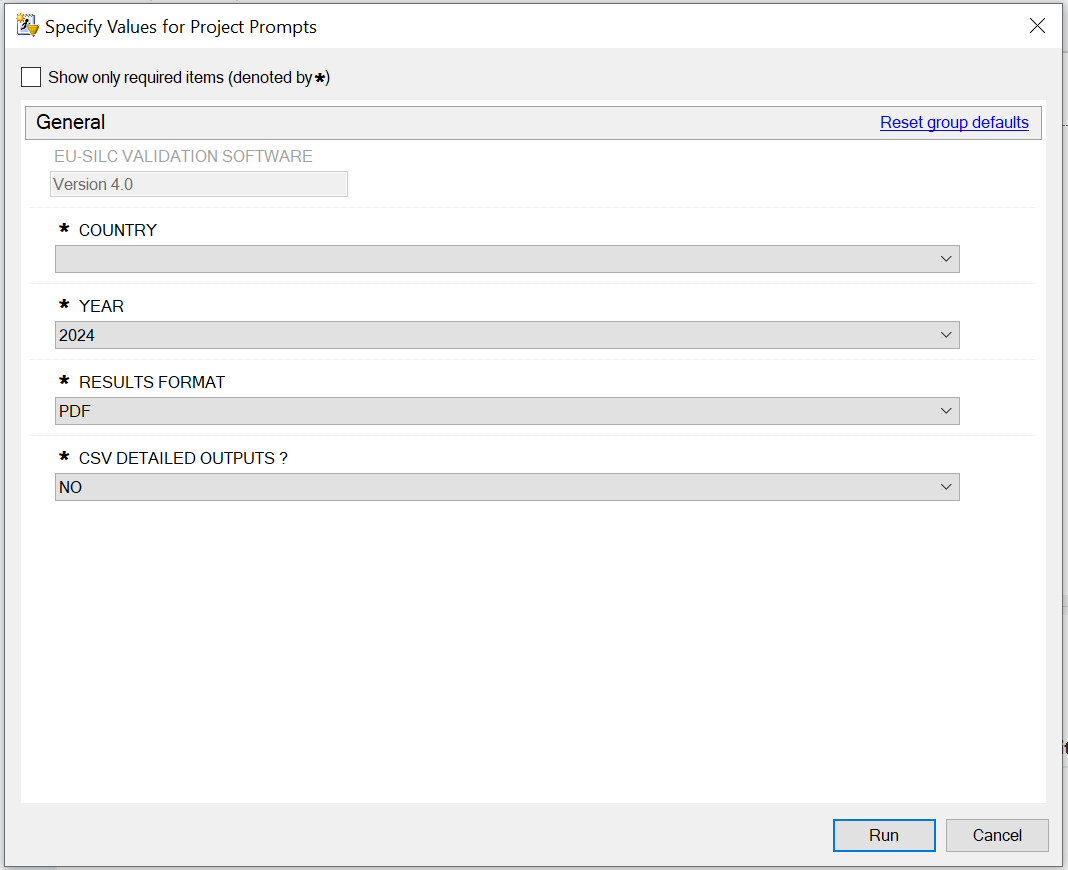
# **SAS EG**

EU-SILC VAL SW is a SAS project made of a number of SAS programs that aim at validating annual EU-SILC data under different perspectives (Fig. 1):

**Figure 1: EU-SILC VAL SW structure**

When running the programme, a dialog box will appear asking the User to specify the execution parameters (Fig. 2).

**Figure 2: EU-SILC VAL SW execution prompt in SAS EG**



The scope of the validation process is defined by the following parameters:

* Country
* Year: it refers to SILC operation year. At the time of writing, 2024 is the default option
* Result options: the User can choose to generate the validation reports in pdf or word format
* CSV detailed options (YES/NO): the User can choose to generate the detailed list of detected errors/warnings for each concerned observation in csv format

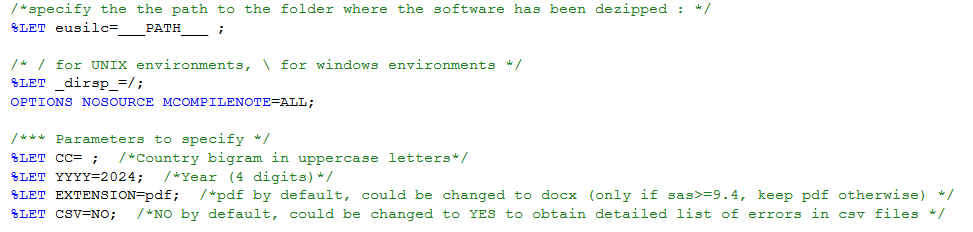
Once the selection is done, click on the “Run” button to start the validation.

# **SAS SA**

Figure 3 shows the execution parameters to set up before running the SILC datafiles validation in case of SAS SA use. These parameters can be set within the **SETENV-XML-SA.sas** module, mentioned in the installation document.

Please store the SILC file to be validated in the inbound directory and check that it is labelled as indicated under section 1.4.

**Figure 3: EU-SILC VAL SW execution prompt in SAS SA**



Once that the execution parameters have been set up, the validation workflow module, i.e.: the main SAS program, can be submitted in the SAS Program Editor window.

The scope of the validation process is defined by the following parameters:

* Country 🡪 *parameter* ***CC***
* Year 🡪 *parameter* ***YYYY***

Once the parameters are set, please launch the workflow running the program “WORKFLOW.sas” located in the **&eusilc/5.3\_Validation/pgm.**

# **The validation reports**

As said, once launched, the VAL SW will automatically generate a number of reports that will be stored in the output directory &eusilc/main/&cc/out where “&cc” represents the country code.

Each generated document’s label will start with ““&CC&YY” where “&YY” indicates the SILC operation year.

The reports produced by the VAL SW can be classified as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| ***Category*** | ***Label*** | ***SILC files checked*** | ***Format*** |
| Overview reports | &CC&YY-Summary | D, H, R, P | pdf/docx |
|  | &CC&YY-Version\_comp |  |  |
|  | | | |
| Detailed reports | &CC&YY-Struct\_Dfile | D | pdf/docx-csv |
| &CC&YY-Struct\_Hfile | H |
| &CC&YY-Struct\_Rfile | R |
| &CC&YY-Struct\_Pfile | P |
|  | | | |
| Analysis reports | &CC&YY-Logical | D, H, R, P | pdf/docx-csv |
| &CC&YY-Samp\_size |
| &CC&YY-Weights |
| &CC&YY-Outliers |
| &CC&YY-Counting |
|  | &CC&YY-Comparison |  |  |

Sections from 3.1 to 3.3 describe the main content of each report. Section 3.4 offers guidance to analyse and interpret the contents of these documents.

# **Overview reports**

* ***&CC&YY-Summary.pdf/docx***

It provides a summary of structural and logical checks performed by the VAL SW. The document is composed of a number of sections, each one related to a different kind of check: missing values count, structural checks, logical checks, etc.

Information is displayed in tabular format by SILC file (each one marked with a different colour) and concerned variables. The table also indicates, for each year, the count/share of the detected warning/error. In case of logical checks, the detailed description of the detected error and the count for each affected year is shown.

* ***&CC&YY-Versions\_comp.pdf/docx***

It is generated to compare the last two versions of SILC data files. To trigger this check, the last two versions of the concerned SILC files should be stored in the inbound directory (&eusilc/main/&cc/csv).

In the heading, it informs on which versions have been checked (that is why it is important to properly number the version of each SILC data file).

For each data file (marked with a different colour), the document reports the differences in terms of number of observations and variables (marked in a green/yellow/red scale according to the severity of the discrepancy) and, in a separate section, it details for each data file the number of values discrepancies detected by variables.

# **Detailed reports**

For the production of these reports, the VAL SW checks several structural aspects: values and flags (with respect to their respective validity domain), household membership status, correct application of routing conditions, correct attribution of non-applicable flags (-8), identification of missing data for calculating the missing value rate.

To produce the reports listed in this subsection, the VAL SW uses EU-SILC methodological guidelines dated until 2020 as reference for data checks.

* ***&CC&YY-Struct\_Dfile.pdf/docx***

This report contains the outcomes of the structural checks carried out on SILC file “D”

* ***&CC&YY-Struct\_Hfile.pdf/docx***

This report contains the outcomes of the structural checks carried out on SILC file “H”

* ***&CC&YY-Struct\_Rfile.pdf/docx***

This report contains the outcomes of the structural checks carried out on SILC file “R”

* ***&CC&YY-Struct\_Pfile.pdf/docx***

This report contains the outcomes of the structural checks carried out on SILC file “P”

For the years prior to 2024 each of the detailed reports will be split in ANTE (with data of years until 2020) and POST reports (with data from 2021 until 2023).

# **3.3 Analysis reports**

The following reports are produced:

* ***&CC&YY-Logical.pdf/docx***

This report contains the result of logical checks carried out between SILC files and, for each detected error, informs on the year, the concerned variables and the affected observations. The heading of each section briefly explains the kind of error.

* ***&CC&YY-Samp\_size.pdf/docx***

This report assesses the achievement of the targets for the following survey elements: sample size for cross-sectional, longitudinal and nuts level, household non-response rate, personal non-response rate, number of proxy interview. For each category, the VAL SW also indicates the grade of feasibility, reporting respective metrics.

* ***&CC&YY-Weights.pdf/docx***

This report displays data on the evolution of rotation groups, summary descriptive statistics for the cross-sectional, longitudinal weights and detected outliers on weight variables. A summary of the main findings appears at the end of the report.

* ***&CC&YY-Outliers.pdf/docx***

This report displays the outliers (observation falling below and above the first and third quartile of the distribution) detected in the SILC files displaying information by variable, year and concerned observations. Each outlier identified can be compared with the lower and upper bound of the variables also reported for prompt reference.

* ***&CC&YY-Comparison.pdf/docx***

This report shows the outcome of the comparative analysis (year to year) carried out on the continuous variables, discrete variables and the associated flags reporting the related variation, showing information by variable and percentile. Descriptive statistics (not applicable, missing, total, negative, positive and zero) are shown for numeric variables both in absolute values and in percentage.

* ***&CC&YY-Counting.pdf/docx***

This report outlines the analysis of nominal and numerical variables, along with their associated flags. For numerical variables descriptive statistics (not applicable, missing, total, filled, negative, positive and zero) are presented in absolute values and in percentages for key categories (missing, negative, zero, positive). For nominal variables the report includes descriptive statistics in absolute terms for the categories (not applicable, total, filled, missing) and the percentage of missing values.

# **3.4 Reports’ analysis and interpretation**

When it comes to analysing the content of the reports, Eurostat suggests to start from the Summary report to have the big picture or the detected error/warnings and to have an idea on the needed actions.

The analysis can be complemented with the detailed report also available in csv format. The csv allows to promptly detect the data point concerned by the identified error/warning.

As shown in Figure 4, the reports in csv format allow to promptly filter the kind of warning (it might affect values, flags, routing conditions, etc.) and to identify the concerned observation.

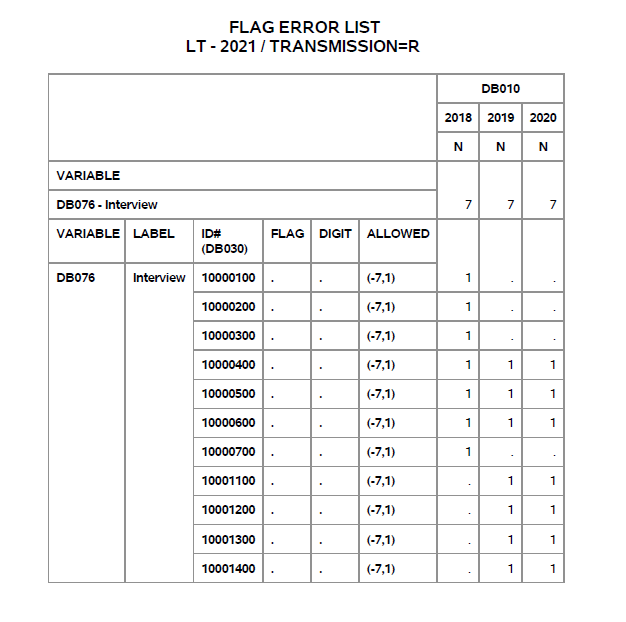
**Fig.4: layout of the csv report**



On the other hand, in the pdf reports, the information is displayed in tabular format by variable and, for each concerned observation (respondent ID), the detail of error for each year – in terms of reported value/flags against expected/allowed value/flags – is shown.

Figure 5 contains an example of the result of the checks carried out on the flags:

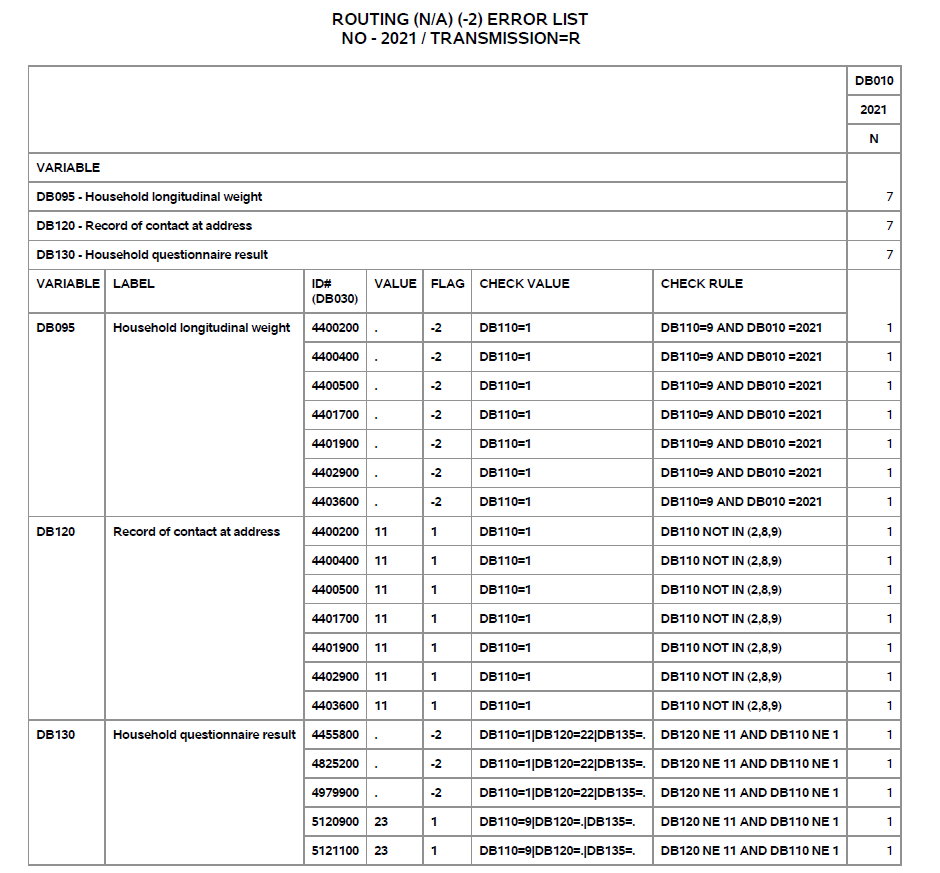
**Fig.5: example of flag error list**



The table should be interpreted as follows: for variable DB076, all the observations under column “ID#(DB030)” did not report any flags (“.” means that no flag was found) for the years indicated in the columns on the right. Note that “1” appears under each concerned year. The table also informs about the expected flags (one of the flags indicated under the “ALLOWED” column – in this case “-7” or “1”).

Figure 6 offers an example of the result of the checks carried out on the routing conditions that aim at verifying if provided data meets the rules foreseen for certain variables in the methodological guidelines:

**Fig.6: example of flag error list**



The table should be interpreted as follows: for variable DB095, DB120 and DB130, reported values/flags did not meet the conditions foreseen in the EU-SILC methodological guidelines. In particular, taking the first error as example, as according to DOC065 for DB095, for the observations listed under “ID#(DB030)”, the reported flag (-2) is not correct as the country filled variable DB110 with 1 (“CHECK VALUE” column) while the condition to have -2 for DB095 is that DB0110 be equal to 9 and DB010 be equal to the last operation year (“CHECK RULE” column).

# **ANNEX: The VAL SW structure**

From an operational point of view, once the VAL SW has been run, the User will not have to perform any action until the validation reports (described in section 3 of this User guide) are generated.

Please consider this Annex as informative as it does not contain any indication of actions to be taken on User’s side.

For the User’s information, as said, the tool consists of a SAS project made of a number of SAS programs (see Fig 1) that treat and analyse different aspects of EU-SILC data:

* XML\_SVAL: it is the core module of the VAL SW. It is built upon a number of macros aiming at parsing the content of the XML file and at checking different aspects related to SILC data file: values and flags (with respect to their respective validity domain), household membership status, correct application of routing conditions, correct attribution of non-applicable flags (-8), identification of missing data for calculating the missing value rate.
* XML\_COLLECT: it retrieves the data from the SILC data files and prepares the SILC files for validation in SAS raw data set format.
* XML\_CHECK: it checks the format compliance between the SILC data files and the XML configuration file of the specific operation year in terms of reported variables/values/flags.
* XML\_INCOME: it has been designed to specifically check if the imputation factor has been correctly reported for income variables[[2]](#footnote-2). In case of mismatch between the imputation factor detected in the data file and the one indicated in the XML file, the SW converts concerned columns in numeric ones but if this operation fails the concerned variables’ imputation factors are reported as non-compliant.
* XML\_DERIVED: it calculates derived variables (on age, working status, etc.).
* XML\_LVAL\_PRE: it provides the functionalities for the logical validation of the derived variables.
* XML\_LVAL: it provides the functionalities (based on SQL queries) for the logical validation.
* XML\_GRID: it checks if the household grid variable has been compiled properly from a structural and semantic point of view. It also carries out some consistency checks that serve as basis for deletion of variables not needed in the process.
* XML\_SAMP\_SIZE: it performs checks on the sample size.
* XML\_OUTL: it detects outliers in the SILC variables and generates the related report.
* XML\_SPLIT: this programme determines, based on the operation year, which validation rules are to be applied to consequently generate ANTE/POST reports.
* XML\_SCL: it checks if variables compiled according to international classifications (geo, NUTS, NACE, ISCO) have been correctly reported[[3]](#footnote-3).
* XML\_SUMMARY: it generates the summary report described in Section 3.
* XML\_WEIGHT: it performs checks on the cross-sectional and longitudinal weights.
* XML\_COMPARISON: it implements the year-to-year comparative analysis on continuous and discrete SILC variables (and accompanying flags) and creates the tables for the relative outcomes in the validation reports.
* XML\_COUNT: it performs the counting analysis of the values in the continuous variables, discrete variables and the corresponding flags.
* XML\_VERSION\_COMP: it compares the last two versions of the SILC data files.

1. Cross-sectional data refer to a given time or a certain time period with variables on income, poverty, social exclusion and other living conditions. Longitudinal data refer to individual/household changes over time, observed periodically over a four-year period (or more years if a longer duration panel is used). [↑](#footnote-ref-1)
2. As known, for income variables, DOC065 allows two/three-digit flags and does not allow missing value. The main source used for the compilation of the variables is indicated by the first digit of the flag, while the second digit indicated the type of collected value. When required, the third digit is used to mark if the value of the variable contains only the kind of components related to the scheme the variable refers to or also other kind of components. In addition, an imputation factor is required: it shows what share of a recorded (transmitted) value is collected and what percentage is imputed. The value should be recorded with two decimal places. Negative values are allowed. [↑](#footnote-ref-2)
3. Cfr. DOC065 Annexes for more information [↑](#footnote-ref-3)