THE LIS USER GUIDE
2024 Template

Table of contents

LIS CROSS-NATIONAL DATA CENTER in Luxembourg

Acronyms

ISO-3166 : The International Standard for country codes and codes for their subdivisions

NUTS : Common classification of territorial units for statistics

ISCED 2011 : International Standard Classification of Education 2011

ISCED 1997 : International Standard Classification of Education 1997

ILO : The International Labour Organization

ICSE-93 : International Classification of Status in Employment 93

ISCO-08 : International Standard Classification of Occupations 2008

ISIC Rev. 4 : International Standard Industrial Classification of all Economic Activities Rev. 4

NACE Rev. 2 : Statistical classification of economic activities in the European Community Rev. 2

CPI : Consumer Price Index

PPP : Purchasing Power Parity

COICOP 1999: Classification of Individual Consumption According to Purpose 1999

Introduction

The *Luxembourg Income Study (LIS) Database* is the largest available income database of harmonised micro-datasets collected from over 50 countries in Europe, North America, Latin America, Africa, Asia, and Australasia spanning over six decades.

The central aspect of LIS' work revolves around the *ex-post* feature of harmonization: LIS acquires pre-existing survey and administrative data in their original shape. Source data entering LIS differ substantially in terms of collection mode (surveys vs. administrative data), type of information collected, level of detail, and structure of the data. Through the harmonization process, the *ex-post* harmonized data are joined together in a common repository, where each set of data (or *dataset*, i.e. a group of individuals and households representing the total population in one year for one country) has the same structure and each variable comparable contents. LIS applies a strict policy to accept LIS data for harmonization. There are various minimum entry conditions:

- The source data must contain income items that ensure a representative and comparable situation of well-being.
- The datasets need to contain the correct micro-level detail, i.e. information needs to be collected at the household and individual level.
- The unit of analysis should be the household; only in rare circumstances, LIS accepts alternative definitions, such as the tax unit, when there is no other dataset available in the country.
- Sampling has to guarantee coverage of the total population, and weighting must ensure that the sample is representative of the total population and subgroups in the country, e.g., by age, sex, regions, or total labor force.

LIS datasets contain *household* and *person* level data on labour income, capital income, pensions, other public social security benefits and private transfers, as well as taxes and contributions and other non-consumption expenditures, and consumption, together with socio-demographic and labour market information. The framework used for the harmonisation of the LIS datasets has evolved over time as a response to changing survey methods, revisions of theoretical foundations of key concepts, and integration of data from new countries. The set of LIS variables, their definitions and their harmonisation rules underlying the current version of the data is referred to as the *2024 Template*.

This document provides generic guidelines for the usage of LIS data, their overall structure and harmonisation practices, as well as an overview of the main blocks of variables included in the *LIS Database*.

More detailed information about the contents and coding of the LIS variables is available in the *METadata Information System (METIS)* under the *Variables definitions*. In addition, METIS provides dataset-specific information in *Contents and Notes section*; this field includes both warnings about deviations from the ideal content or informative country specific variable notes, as well as the content of income and consumption variables in each dataset. Information about the original data source, such as the data provider, data collection information, and additional technical information is also available in *METIS* in the *Dataset information* section. In addition, *Compare.It* provides a country-specific summary of data comparability-related issues.

THE LIS USER GUIDE
2024 Template

I. Overall Structure and generic rules and practices

1. Naming of the datasets

The datasets in the LIS database are named according to the 2-character country abbreviation coded according to the *ISO-3166* and the income reference year.

The income reference year is ideally the calendar year for which income data has been collected. If the income reference period crosses over two years, the year with the longest period will be chosen; if the reference period is equally split between two years, the most recent year will be used. Please note that the income reference year may differ from the year by which the data provider identifies the dataset (e.g., for all datasets that used *SILC* database as input, LIS year is *SILC* year-1).

2. Structure of the datasets and sample selection

Each LIS dataset is composed of two files, a household level file and an individual level file.

The **household level file** (henceforth referred to as the LIS **H-file**) contains a record for each household of the sample. The data provider defines who composes a household. The most common definition of household used by most data providers is the single person or the group of persons living in one dwelling and sharing a budget). Note that in a handful of cases the data are based on concepts different than the household (e.g., tax unit, family unit, etc. – see *Dataset information* in METIS for more information).

The **individual level file** (henceforth referred to as the LIS **P-file**) includes as many records as there are individuals in the household.

In all datasets, the household and individual level files contain the same variables. This means that all variables are physically present in the file even if the information was not available for that country in that year.

Sample selection and household membership

The sample included in the LIS P-files represents a cross-section of the total population during a specific time period.

Household sample - The final sample included in the LIS H-file includes only those household-level observations belonging to the representative cross-section of the population selected by the data provider according to their selection procedure.

Individual sample - The LIS P-file includes observations for all *household members,* i. e. individuals that belong to the households included in the LIS H-file and that share the budget in the household.

Note that in some surveys, information is collected also for individuals who are linked to the household but do not share the budget (this applies to live-in domestic servants, lodgers, guests and boarders, and in some rare case also to persons not living in the same dwelling, such as relatives living temporary elsewhere for work or study). In these cases, LIS discards them from the final LIS sample because they are not considered household members according to LIS definition.

Exceptionally, LIS keeps these non-household members in the sample only if their incomes were aggregated in the household incomes by the data provider without the possibility to be separated from the household members' income.

3. Variable standardisation

Standardised variables

The majority of the variables in the LIS template are *standardised*, this means harmonised along two dimensions: in terms of conceptual content (the variables aim to capture the same concepts across datasets and are as comparable as possible in terms of concepts/ definitions) and in terms of coding structure, i.e. that LIS systematically groups the national information to more broadly defined categories.

There are two blocks of standardised variables. Standardised *continuous* variables report information expressed in the same unit across different datasets (e.g., monetary variables that are components of the total income report annual units of national current currency, hours variables report number of hours worked per week, age variables report number of years). Standardised *categorical* variables report information expressed with the same value codes and labels.

Country-specific variables

There are several variables that are not standardised (variables with *"_c*" suffix) because, even if they aim to capture the same concepts, the values are country specific. An example of such variable is *region_c*: administrative regions, even if it is expressed in a standard classification such as *NUTS*, the same values have a different meaning across countries. This is indicated by a differing label documenting the actual (dataset specific) content for the dataset in question. Thus note, both the exact content and the coding structure will differ across datasets. Even within the same country series, these variable can differ from one year of another (for example codes of *region_c* if an administrative reform took place or there are less or more details available in the original data in different years) The *country-specific* (*_c*) variables are always *categorical* and their exact content and coding structure is indicated in the (dataset-specific) variable and value labels (as can be seen in the *Statistics* or *Codebooks* sections of *METIS*), while additional information may be provided in *Compare.It.*

4. Generic missing values policy

In the original data used as a source for the LIS Database, the treatment and documentation of the missing values (refused to answer or does not know) and of the not applicable cases, vary substantially. In order to simplify the information and usability across the harmonized datasets, LIS has standardised its approach to missing values in the following way.

Missing values policy for income and consumption variables - This section applies to the blocks Current income, Income deductions, transfers paid and loan repayments, Extraordinary income, Imputed rent, Consumption expenditure, and Major economic aggregates. LIS codes missing values to value ‘0’ in the following two scenarios:

i) the information has not been collected;

ii) the information has been collected, but is not available at the level of detail necessary for the LIS variable in question.

Only observations where respondents refused to answer or indicated that they do not know the amount, are reported with missing value ‘dot’.

Missing values policy for other LIS variables - The missing value ‘dot’ generally indicates observations where information is not available. This includes cases where the information is not applicable, for example when a person does not work and therefore has no information on current occupation (section *Characteristics of Main Job* remains all missing for this person). It also includes cases where the information is applicable but missing. In these instances, the information may not have been collected by the data provider for a specific subset of the sample, or at all (for example, only reference persons and their partners were asked certain questions). Additionally, the information might have been collected, but the respondent did not provide an answer, such as in cases of "don't know" or refusal.

5. Weighting

Survey weights are created to correct coverage, sampling, and non-sampling errors, including biases from unit and item non-response. Their purpose is to ensure that, while used, the sample accurately represents the entire target population. Since an unweighted sample reflects only itself and not the broader population, calibrated weighting factors, which align sample estimates with known population totals, are essential for statistical analyses if a researcher intends to make valid inferences about the target population or specific subpopulations.

All *LIS* files include weight variables calculated by the data producers (at the household level in the H-file and at the individual level in the P-file) that are needed to make the sample representative of the overall population. All of them correct for errors and biases, but do not necessarily inflate to the covered population (please refer to *Data Information* in METIS for more information on weights). Suppose the weight initially provided by the data producer does not inflate to the covered population. In such a few cases, an inflation of the weight to the covered population is conducted by LIS team using the following formula:

$$w_i' = w_i \frac{p}{\sum_{i=1}^{n} w_i}$$

where w_i' is the adjusted weight for each observation so the total matches the given population size, w_i is the original weight for each observation, p is the total population size, and $\sum_{i=1}^{n} w_i$ is the sum of the original weights across all observations. This formula rescales each weight proportionally, ensuring consistency with the population while preserving the relative contribution of each observation.

The variables *hpopwgt* and *ppopwgt* are household and person-level weights normalized to population size and available for all *LIS* datasets. Note that when only the household-level weight was provided, the individual-level weight is set equal to the household-level weight for all household members.

Suppose a researcher pools (appends) datasets together for cross-country analysis. Among many, there are two countries whose populations differ significantly, e.g., Luxembourg and Brazil. Using the weights that inflate to the population size of each country (*hpopwgt* or *ppopwgt*) in the regression analysis would make the results biased toward Brazil because of the Brazilian population size. Therefore, the weights need to be normalized so a country with a larger population size (Brazil) does not take over the results from a country with a smaller population size (Luxembourg). One way to normalize survey weights across different samples for cross-country analysis (adopted by LIS team) is to adjust the weights so they sum up to 1 by dividing each observation's weight by the total sum of weights, ensuring proportionality across observations:

$$w' = \left(\frac{w_i}{\sum_{i=1} w_i}\right) \cdot 10{,}000$$

Scaling the weights with multiplication by 10,000 is done for practical use in statistical analyses while maintaining relative proportions. The normalized weights are scaled so that their sum is approximately 10,000. Normalized and scaled weights as described above are represented in variables *hwgt* and *pwgt*, which represent relative proportions within each country and preserve the **relative importance** of observations within a dataset, independent of the country's population size or sample size. Normalizing and scaling weights ensures that each country's dataset contributes proportionally in both descriptive and inferential analyses.

Finally, LIS also provides two additional weight variables (*hwgta* and *pwgta*) if some questions in the original survey were asked only to a subsample of individuals/ households, and subsample weights were provided in the original data.

6. *Gross, net* and *mixed* income datasets

LIS datasets are classified into either *gross*, *net* or *mixed* income datasets depending on the extent to which income taxes and social security contributions are captured.

Gross income datasets

In the ideal case, all LIS income variables (with the exception of the disposable household income variables, *dhi* and *dhci*) report amounts gross of income taxes and social security employee contributions, and the overall amount of taxes and contributions is also available separately, so that it can be deducted from the total gross amount in order to obtain total disposable income. All the datasets satisfying these criteria are referred to as "gross" datasets (see variable *grossnet* – codes

100 to 120). Please note that often taxes and contributions are only available at the household level even for datasets that have individual level incomes because the family is the tax unit in many cases.

Net income datasets

Often, however, respondents are asked to report only net amounts (as those are the ones they know best), and there is no information about the income taxes and social security contributions paid on those amounts. In these cases, all LIS income variables report net amounts, including the overall total income variable, which will thus be the same as disposable income. These datasets are referred to as net datasets (see variable *grossnet* –code 200).

Mixed income datasets

There are some datasets that are neither fully net nor gross. This can happen when information was available only for taxes but not for contributions (or vice versa), or when the information on taxes and contributions was only available for certain subcomponents of total income. Depending on the specific situation, the LIS income variables may either be all net, or all gross, or gross of only taxes or contributions, or partly gross and partly net (a dataset note in METIS will inform the user about the specific situation). All those cases are flagged as being “mixed” income datasets (see variable *grossnet* – codes 300 to 320).

Please note that the term gross and net as explained above should not be confused with the situations where the same terms refer to incomes that are gross or net of costs (e.g., rental income and self-employment income can be recorded either before or after deduction of the costs incurred); with respect to this definition of gross/net, LIS variables are net of the costs (and a definitional deviation note would warn users when this is not the case).

7. Annualisation

All LIS income flows and consumption variables report annual amounts. If the original survey did not provide annual amounts, whether because of a different reference period, or because the amounts are collected as usual amount together with periodicity (e.g., usual monthly wage, and number of months during which it was received during the year), LIS annualises the amounts. In the case of current income surveys (where the incomes refer to the last payment received), or surveys with shorter reference period than a year for the incomes/ expenditures (this is often the case for household budget surveys, which may collect for example all the inflows and outflows during a given month), the annualisation carried out by LIS involves the assumption that those flows were occurring during the whole year with the same pattern as during the reference period (and the reported amounts are multiplied by 52 if weekly, 12 if monthly, 4 if quarterly, etc.).

Lump-sum incomes are considered as a unique payment therefore they are not multiplied by any number unless there is a clear reason to do so (e.g., the original survey indicates that the lump-sum income is received twice per year).

8. Aggregation rules

Aggregation of sub-categories into overall categories in categorical LIS variables

Many categorical variables have a nested coding structure whereby the multidigit codes starting with the same digit belong to the same upper category. As a result, by selecting only the subcategories, one does not necessarily get the total of one overall category. For example, in variable ***own*** (main residence tenure status) the codes in the 100s stand for *owned housing*, while the codes 110 and 120 are subcategories of the former upper-level category, namely *owned outright* and *owned with mortgage*. In one and the same dataset observations can be coded either at the higher level or at the lower level (e.g., some households may have been coded 110 (*owned outright*), others 120 (*owned with mortgage*), yet others directly with 100 (*owned*) in case it was not possible to determine whether they have a mortgage or not). As a result, by selecting all the subcategories, one does not necessarily get the total of overall owned category (e.g., by selecting *owned outright* (110) and *owned with mortgage* (120), one does not necessarily select all classified as homeowners; the selection of the higher code 100 is needed as well).

Aggregation of individual level incomes into household level incomes

In general, household income amounts are aggregated from individual-level values of all household members, so that the sum of any LIS individual income variable in the P-file over all household members is identical to the amount in the corresponding LIS household level variable. In some cases though, individual level incomes do not sum up to household income; this may be because:

i) only a part of those incomes were collected at the individual level, while the rest was only available at the household level (e.g., the income received by children under the minimum age to answer to the individual questionnaire is recorded only at the household level). If individual-level information is not available, all the individual level income variables remain at zero and the recorded household level information is used to fill in the household level LIS variables directly;

ii) in rare circumstances, the individual level income variables and the household level ones are filled independently from each other (this happens if the data provider provides two sets of variables, either because they were collected independently, or because only in one of the two sets some variables (e.g., taxes) were imputed, e.g., cn13).

Aggregation of LIS income variables

LIS *total current income* (*hitotal*) is the sum of five major sources: income from *labour*, *capital*, *pensions*, *other public social benefits* and *private transfers*. Each of these five main income blocks is aggregated from further subcomponents. While the aggregation of total current income from the five main income blocks is always ensured (i.e. *hitotal* = *hilabour* + *hicapital* + *hipension* + *hipubsoc* + *hiprivate*), the further breakdown of each of these five blocks into subcomponents is not always available due to data limitations. Some amounts are reported included conceptually at the higher level. For example, the variable list only contains *hi21* (interest and dividends) and *hi22* (rental income). Other capital income such as royalties and other capital income from investment in self-employment activity are not reported separately in the LIS data. Instead, they are directly added to *hicapital*. Hence, the formula for constructing the major block capital income is: *hicapital = hi21 + hi22* + additional amounts directly placed in *hicapital*.

Likewise, in datasets with very little level of detail available, for example when total labour income was provided as a unique amount, then this amount is directly put into *hilabour* and in that case the four subcomponents are coded all with value zero.

The same aggregation rationale applies to the individual level income variables: *pitotal* is always equal to the sum of all individual level variables (*pilabour*, *pipension*, *pi411*, *pi42*, *pi43*, *pi44,* and *pi511*), whereas the subcomponents of each of them do not necessarily add up to the upper-level aggregate. It should be noted that variable *pitotal* is always created, even when not all the subcomponents are available at the individual level (and especially it is the case for certain social benefits); as consequence, it cannot be ensured that *pitotal* is comparable across datasets.

Note that the sum of *pitotal* of all household members **is always lower than** *hitotal* because of all the incomes provided only at the household level (e.g., social benefits such as child allowances, general assistance, housing benefits, and capital income).

For the blocks of extraordinary income, no upper level is provided and the same applies for income deductions, transfers paid and loans paid block, the subcomponents being too heterogenous.

Aggregation of LIS consumption variables

Consumption information is available in a much less harmonised way. Whereas some datasets are created from very detailed household budget surveys, other surveys include only a very limited set of consumption items. LIS only creates the sum over the subcategories when there is sufficient information to analyse household consumption expenditure (LIS variable *hcexp*). In case there is only a subset of consumption items available *hcexp* contains all 0s. It is advised to consult *METIS* and *Compare.It* to screen for information about consumption data before analysing consumption data.

9. Data editing and imputation

Consistency checks

When constructing the LIS variables, LIS runs automatised consistency checks of the original data, and corrects, if necessary, specifically for the living arrangements variables. For example, it is always ensured that each household has only one reference person.

Note that LIS deliberately refrains from editing outliers in its microdata, as these extreme values can provide important insights into the full distribution of the data. By leaving such values in place, the dataset maintains its integrity, reflecting the real diversity and range of responses, which could be crucial for understanding broader trends and patterns.

Imputation

Depending on the data provider's practices, some LIS datasets include missing values for item non-response, while others replace missing values with imputed data. In cases where imputations were performed by the data providers, LIS generally uses the imputed values, but they are not identifiable in the LIS data.

In rare cases, when the data provider does not perform imputations and a large number of missing values could bias the analyses, or when taxes and social contributions are either not fully available or completely missing, LIS carries out imputations in-house or contracts experts to simulate the missing taxes and contributions.

II. **LIS Variable Groups**

1. Income and Consumption variables-LIS flow variables

LIS datasets include a large number of variables referring to incomes (current and extraordinary income), non-consumption expenditures and consumption. Data production for such variables may differ substantially between data providers: some data are based on surveys, others on administrative records, or a combination of the two.

All LIS income and consumption variables are reported in annual amounts and in units of the national currency in current use (see variable *currency*). In addition to this, LIS provides a dataset that includes conversion factors for CPI adjustment and PPP adjustment to 2017 and 2011 International Dollars. This dataset can be accessed online (https://www.lisdatacenter.org/resources/ppp-deflators/) and through LISSY (see https://www.lisdatacenter.org/data-access/lissy/syntax/).

LIS income and consumption variables are split into the five major blocks:

- ***Current income*** (variables prefixed by *hi* and *pi*): this consists of monetary payments as well as the value of goods and services received by the household, typically at regular intervals (e.g., monthly). In certain cases, one-time payments such as birth grants, sickness/injury benefits, or specific labour-related supplements like bonuses may also be included, provided they are intended for current consumption and that their spending does not reduce the net worth of the household;

- ***Income deductions, transfers paid and loans repayments*** *(*variables prefixed by *hx* and *px*): these consist of non-consumption expenditures such as taxes, contributions, inter-household transfers and loans paid, including mortgage (incl. interests);

- ***Extraordinary income*** (variables prefixed by *he* and *pe*): this consists of windfall gains and other irregular and typically one-time receipts;

- ***Imputed rent***: this block consists only of the imputed rent *(variable hrenti)* which is not added up to any of the income or consumption blocks, it is up to the users to add it if they consider it part of income/ consumption;

- ***Consumption*** expenditure (variables prefixed by *hc*): these consist of monetary and non-monetary consumption items, following COICOP 1999 classification, with the exception of imputed rent that is not part of this block.

Additionally, LIS provides several ***major economic aggregates*** constructed from variables belonging to the above-mentioned major blocks and consisting of a series of income aggregates (including the overall concept of ***total disposable income*** and its major components) and consumption aggregates (e.g., housing costs).

While all these variables exist at the household level in the household file (prefixed by *h*), however, only a part of them are created at the individual level in the person file (prefixed by *p*), depending on the relevance of the variable at the individual level (e.g., housing benefits, see *LIS Variables List* for an exhaustive list of the variables that exist in the person file).

1.1 Current income

Current income consists of monetary payments as well as the value of goods and services received by the household, typically at regular intervals (e.g., monthly). In certain cases, one-time payments such as birth grants, sickness/injury benefits, or specific labour-related supplements like bonuses may also be included, provided they are intended for current consumption and that their spending does not reduce the net worth of the household. These include monetary and in-kind income from labour, income from capital, pensions, other social security transfers, and in-kind social assistance transfers, as well as monetary and in-kind private transfers. There are two notable exceptions among the non-cash incomes:

1. *Income from household production of services for own consumption*

   These refer to the *value of services from household consumer durables*, the *value of unpaid domestic services*, and *the net value of owner-occupied housing services*. Especially the first two are rarely available in the microdata and, when available, they are often calculated using different methodologies. Thus, the *value of services from household consumer durables* and the *value of unpaid domestic services* are not included in the LIS database. The *net value of owner-occupied housing services* (imputed rent) is excluded from the calculation of LIS total income (*hitotal*) and disposable household income (*dhi*), a variable for imputed rent (*hrenti*) is provided in a separate section, when available.

2. *Social transfers in kind (STIK) received*

   These refer to government-provided services that benefit individuals, but are provided with the primary objective of meeting the general needs of the overall population, rather than that of assisting the poor. Specifically, LIS does not include in-kind transfers in the areas of health care and education. The value of these in-kind services is very difficult to estimate at the individual/household level and national ways how to allocate the benefits may vary substantially. In addition, these values have rarely been calculated in the surveys. For this reason, the value of these transfers is also excluded from the LIS microdata.

It should be noted that LIS follows with the latter broadly the *Canberra Group Handbook on Household Income Statistics* which generally excludes social transfers in kind (STIK) from its operational definition of disposable income due to their valuation challenges and non-availability. Note however, when provided by the data provider LIS would include public housing subsidies. On the other side, although the Canberra group includes the net value of owner-occupied housing services in their operational definition, LIS keeps these values excluded from the disposable household income (*dhi*) due to their differences in national valuation and non-availability in various countries, which would limit cross-national comparability.

However, LIS acknowledges the importance of STIK in capturing broader aspects of household well-being, particularly in extended income measures, where their estimated value—such as imputed rent for housing benefits—is included to provide a more comprehensive view of economic resources.

Total current income is disaggregated into five major components:

Current income from labour (*h/pilabour* and detailed variables with prefix *h/pi1*): monetary payments and value of goods and services received from dependent employment, as well as profits or losses from self-employment (including in-kind) and value of goods and services produced for own consumption.

Current income from capital (*hicapital* and detailed variables with prefix *hi2*): monetary profit from property and capital (both financial and non-financial assets), including interest and dividends, rental income and royalties, and other capital income from investments.

Pensions (*h/pipension* and detailed variables with prefix *h/pi3*): pension income from all three pension pillars: public, occupational, and private pensions. The public pensions are further distinguished by type: insurance, universal, and assistance.

Income from public social benefits excl. pensions (*hipubsoc* and detailed variables with prefix *h/pi4*): social security transfers (excluding public pensions) stemming from insurance, universal or assistance schemes, and in-kind social assistance transfers (most common being food).

Private transfers (*hiprivate* and detailed variables with prefix *h/pi5*): monetary transfers and the value of in-kind goods and received from private institutions and other households. The only exception are the merit-based scholarships that can be from a state educational institution as well.

1.2 Income deductions, transfers paid and loans repayments

These consist of non-consumption expenditures such as taxes, contributions, inter-household transfers and loans instalments.

1.3 Extraordinary incomes

Extraordinary income consists of all windfall gains and other such irregular and typically onetime receipts, such as windfall labour income (severance pay, retirement packages offered by the employer), capital gains, inheritances, and other extraordinary income (insurance compensations, lottery winnings, etc.).

1.4 Imputed rent

This block consists only in the imputed rent variable, defined as the imputed value of the yearly use of the dwelling used as a main residence owned by the household or occupied rent free or at a reduce rent (and in that case the imputed rent is only the difference from the market rent and the actual rent paid).

1.5 Consumption expenditures

LIS ideally records total consumption, including that stemming from expenditures (i.e. if the good or service consumed has been purchased by the household) and that stemming from own-production, transfers or gifts (the value of goods and services consumed which were given to the household by somebody else, or self-produced).

Consumption items (whether monetary or non-monetary) can be disaggregated by the type of good or service being consumed (LIS uses the 12 major groups of the *COICOP* 1999 classification of consumption goods and services except the imputed rent which is not included, and reported in a separate section due to data limitations).

1.6 Major economic aggregates:

In order to facilitate the use of the LIS data, LIS has created some income and consumption aggregates that are commonly used in research. These aggregates are either derived from other LIS variables, in which case the calculation always follows the same formula to ensure perfect comparability across datasets, or are produced from original variables (in which case it is not possible to recreate them on the basis of other detailed LIS variables – this is true for the breakdown of public transfers in its three categories insurance, universal and assistance transfers). Are listed below those aggregates that are not part of the *current income* aggregation:

- *dhi*: disposable household income, which includes current total income net of income taxes and social security contributions (*dhi = hitotal - hxitsc* for gross datasets, for net datasets *dhi = hitotal*, therefore it is the variable used to compare across gross and net datasets). *Dhi* is also the variable used for the LIS Inequality and Poverty Key Figures.

- *hvalgs*: total value of goods and services, which includes the in-kind income from labour (fringe benefits and own consumption), in-kind transfers from the State, as well as from private institutions and other households (*hvalgs = hi13 + hi14 + hi47 + hi53*).

- *dhci*: disposable household cash income, which includes total current monetary income net of income taxes and social security contributions (*dhci = dhi – hvalgs).*

- *hpublic*: public transfers, which corresponds to the social security redistribution, and includes public pensions and other public social benefits; this variable is further disaggregated into:

  - *hpub_i*: insurance transfers: transfers stemming from social security systems where eligibility is based on the payments the of the social security contributions and the duration of such payments

- *hpub_u*: universal transfers: transfers stemming from public programmes that provide flat-rate benefits to all residents, or only a certain category of residents, provided that they are in a certain situation, without consideration of income, employment or assets; note that in some cases the benefit amount may also depend on the other incomes of the individuals, which may result in some proportion of the population at the upper end of the income distribution being excluded from receipt;

- *hpub_a*: assistance transfers: transfers stemming from public programmes that provide benefits targeted to individuals or households in need, subject of an income and/or assets test; the amount of the benefit is either a flat rate or is based on the difference between the recipient income and a standard amount representing the minimum subsistence needs as guaranteed by the government.

2. Other variables

2.1 Technical variables

Technical variables consist of file identifiers, dataset level information, weights and imputation flags.

File identifiers and dataset-level information - Each file contains unique identifiers of its observations (*hid* for the H-level file and both *hid* and *pid* for the P-level file). In addition, in both files there is a series of dataset-level variables that provide information about the dataset and that are useful when working with multiple datasets (*did, dname, name, iso2, iso3, country, wave, year, currency* and *grossnet*).

Weights – Each file contains the weights that inflate to total population (*hpopwgt* and *ppopwgt*) and the normalised weights (*hwgt* and *pwgt*). In addition, variables *hwgta* and *pwgta* report additional household/individual level weight in case only part of the household sample has been selected for some variables and the data provider calculated a special weight for that subsample.

2.2 Geography and Housing variables

These variables are all household level variables and report geographic as well as dwelling information.

Geographical characteristics contain country-specific information about the locality in which the household resides, the geographical/administrative region, as well as indications about the population density or type of area. LIS has created a dichotomous indicator for rural areas (*rural*) which, in datasets that do not have direct information on rural, is constructed based on country specific information on type of area.

Dwelling characteristics contain information about the dwelling that is the principal residence of the household: the tenure (owned versus not owned), the type of dwelling (house versus multi-unit residential building or other type of dwelling) and number of rooms.

2.3 Household composition and living arrangements

These variables contain household level information about the composition of the household and individual level information about living arrangements of household members.

Household composition variables are generally derived from individual level demographic and living arrangement variables.

Living arrangement variables include the relationship of each household member to the reference person of the household, as well as indications on whether the household members have a cohabiting partner, or whether they live in the same household as their partner, parents or children. Please note that the *household reference person* is usually chosen by the data provider, and its definition differ across datasets (main income earner, person most knowledgeable about the budgetary situation of the household, eldest person, person responsible for the dwelling contract, or simply self-defined by the respondents, etc.). Whereas the relationship to the household reference person is typically available for all household members (with some major exception for the data from early years), the presence of a partner, parents or children is often available only for household reference person (and implicitly his/her partner when there is one).

2.4 Socio-demographic variables

These variables are all individual level variables and report the major socio-demographic characteristics of the household members. They are split into four major blocks: demographic characteristics, immigration, health and education.

a) Demographic characteristics

These refer to the age, gender and marital status (both *de facto* and *de jure*) of the household members. *De jure* marital status is reported in the categorical variable *marital*, while *de facto* status (the cohabiting situation) in the dichotomous variable *partner* (and *hpartner* for the reference person).

Age and gender are always available for the totality of the sample (with the exception of some earlier datasets for which individual characteristics were only collected for reference person and spouse, or only for adults, or up to a certain number of adults). In rare cases, age is provided in intervals.

b) Immigration

These variables are intended to identify the immigrant status in a given country. They include information on citizenship, country of birth, duration of residence in the country, ethnicity and other relevant country-specific immigration characteristics (such as mother tongue, religion, etc. that could be the content of *immigr_c* or internal migration). LIS creates a dichotomous variable (*immigr*) on the basis of the available information on citizenship and country of birth (see the definition of the *immigr* variable in METIS for a precise indication of the rules used for its construction).

c) Health

Indication of a disability above a certain level or other information on the health status has been used to create a dichotomous variable (*disabled*) to identify persons suffering of a certain disability or a chronic disease that limits them in their daily activities. Additional information on health is provided in the country specific variable *health_c* that includes mostly subjective health status measured on a scale from poor to excellent (however be aware that the scale can differ across datasets).

d) Education

Education variables include information about the highest education level achieved by the person, total years of education, as well as an indication of whether the person is currently enrolled in education and some additional education-related information (such as parents' education level).

The *highest education level* as provided by the data provider is reported in the variable *educ_c*; ideally, this refers to the highest completed level, but in some cases, it might refer to the highest attended level. The variable *educlev* contains the highest completed education level harmonised according to the 8 major categories of *ISCED* 2011 (or *ISCED* 1997 for older datasets). Finally, the variable *educ* provides a 3 categories classification of the highest completed education, where low stands for less than upper secondary (corresponding to *ISCED* 2011 levels 0, 1 or 2), medium for upper secondary and postsecondary non-tertiary level (*ISCED* levels 3 or 4) and high for tertiary education (*ISCED* levels 5 to 8).

In case it was not collected as such in the data, the number of education years (*edyrs*) has been derived from the highest completed level according to the average duration of each cycle in a first stage (see variable definition in *METIS* for more details). The *illiterate* dummy flags individuals who cannot read and write in any language. Country-specific information about the education level of the mother and the father of the household members is also provided when information about the education of the parents not living in the same household is available in the original data.

2.5 Labour market information

These variables report information on the labour market participation of the household members (individual level variables with the exception of *farming* which is recorded at the household level). They are split into two major blocks: labour market activity and characteristics of the main job.

a) Labour market activity

This section reports variables about the labour market activity current status, as well as information about employment intensity during the reference year and overall years of work experience or indication of employment in the informal sector.

Labour activity status is reported in the labour force status variable (*lfs*), which aims at capturing the main current activity status. If the current main activity status is not available, the main activity status in the income reference time will be used. In case the latter is not available either, the employment status according to *ILO* criteria in current period will be used instead (dataset-specific notes in *METIS* inform about the exact situation in each dataset). The labour force status distinguishes between the employed (those for whom work is the main activity during the current period), unemployed (job seekers) and not in labour force (inactives). Among those not in labour force the variable distinguishes between those retired from a job or business, those unable to work due to long-standing health problems, those enrolled in education and those inactive because they are fulfilling domestic tasks. The concept of main activity is generally defined in the original data (typically it will be self-assessed by the respondent, following some instructions such as "the activity at which you spend the most time").

The dummy variable *emp* provides an employment flag that is existing in all datasets and that is mostly consistent with the job characteristics variables. On the other hand, the dummy *emp_ilo* (which flags individuals employed according to the *ILO* criteria) is filled only when the information was available in the original data. Are considered *ILO* employed persons who worked for at least one hour for pay or profit in a short reference period prior to the survey or had a job but did not work in the short reference period due to temporary absence from the job because of sickness, maternity leave, holidays, etc. or due the nature of their working time arrangement, such as shift work, etc.

Please note that activity status variables are typically defined only for those persons who answered the labour market individual questionnaire (adult persons for most datasets). The only household level variable of this section (*farming*) flags the households that are carrying out farming activity (crops and/or livestock).

Of particular importance for the lower and middle income countries, the *informal* dummy aims to capture information about the involvement in the informal sector. For employees, this could refer to a situation where they work without a working contract or they do not contribute to the social security system, they work in an unregistered business, they do not beneficiate of legal rights (right to pension, paid leave, etc.) or their wage is under-declared. For the self-employed, this could refer to a situation where they own an unregistered business when the legislation in the country requires them to register it, they do not pay taxes and/or contributions if they have to pay them. Those who produce goods and services only for their own consumption are out of the scope of the definition of informality of this variable.

Several measures of overall employment intensity include indicators of multiple jobs holder (*secjob*), the number of hours worked per week (*hourstot*), number of weeks worked in a year (*weeks*) and full-time weeks (*weeksft*). On the basis of the number of full-time weeks worked (or overall weeks and the current hours worked if the former information is not available), LIS creates an overall full-year full-time dichotomous variable (*fyft*). There is also a dummy that flags parents in maternity/adoption or parental leave (*parleave*).

Please note that different employment intensity variables are defined for different groups of the population: variables about labour force status, hours worked, multiple job indicator and informal activity refer to the present situation, and are hence defined only for persons who currently work (as from the *emp* dummy variable); on the other hand, variables about weeks worked during the year and full-year full-time indicator, refer to the situation over a longer reference period (usually the income year), and are hence defined for all persons who were eligible to answer the labour market individual questionnaire (usually all adults), whether they are currently working or not. As a result, for a person who is not currently employed weekly hours will not be reported (the variable will be set to missing), whereas annual weeks will be reported (at either zero if the person did not work during the whole reference year, or any positive number if the person worked at some point during the year).

Finally, the variable *wexptl* refers to the overall work experience, with indication of the existence of such experience, as well as its duration (total years of total work experience). This variable is typically defined for the same group of persons for which the labour status variable is defined (usually all adults).

b) Characteristics of the main job

These variables report the most important job characteristics for the main job of every employed individual. They include the status in employment (dependent employee versus self-employment), the industry, the occupation, the sector of employment (private versus public), type of contract (permanent/long-term versus short-term contract), the work intensity (weekly hours worked in main job, and a part-time dummy), and an indication of the income level (as indicated by the monthly wage and hourly wage rate).

Job characteristics variables are standardised ones, except one variable for occupation (*occ1_c)* and one for industry (*ind1_c*) and which report the respective information (in most cases an international classification) at the most detailed level available in the original data.

Job characteristics variables for the main job are defined only for persons who currently have a job /typically those employed in *emp*).