

SOUTHMOD Modelling Conventions

EUROMOD Modelling Conventions adjusted for tax-benefit microsimulation in developing countries

Last updated: 7 March 2024

This document provides the SOUTHMOD modelling conventions. Initially based on the EUROMOD modelling conventions, these modelling conventions have been amended and extended for use in a developing country context relevant for the collaboration between UNU-WIDER, ISER, SASPRI and national teams of the respective countries in the SOUTHMOD project.

The guidance put forward in this document supports the process of building a model and its subsequent upkeep. It is shared publicly as a courtesy for any researcher who aspires to build his/her own model based on the freely available EUROMOD/SOUTHMOD software as part of the SOUTHMOD project or independently. For the quality of any model built in this context it is crucial to carefully and systematically check the data going into the model and all elements of the model.

This document strives to streamline conventions in terms of data and modelling across countries that are part of the SOUTHMOD project. Nevertheless it is ultimately the responsibility of each person using the models to understand and make a careful decision regarding the extent of comparability of the models when applying them to his/her research question. We therefore strongly recommend that the country report and associated Data Requirement Document (DRD, available from UNU-WIDER upon request) are studied carefully and, if there are further questions, that UNU-WIDER is contacted for the Stata do-files creating the data which underpins the SOUTHMOD model(s).

Where appropriate, each section of the guidelines is divided into two parts: essential (compulsory) and desirable (i.e. not essential but may improve the model if possible to implement). In some sections all the modelling conventions are categorised as essential.

Contents

1	General	1
2	Input datasets	3
3	Variable naming convention	13
4	Uprating factors.....	14
5	System and database configuration.....	15
6	Scope of policies.....	16
7	Policies.....	18
8	Switchable policies (extensions).....	20
9	Tax units	22
10	Income and expenditure concepts and relationship with income lists.....	23
11	Income lists	25
12	Indirect taxes.....	29
13	Output	31
14	Validation	32

1 General

Essential

The term “**policy year**” or “**policy system**” refers to the tax and benefit policy rules as of 30 June or 1st July in the given year. See also Section 6.

The term “**income/expenditure reference period**” implies the time period to which the income and expenditure information in the input data refer (e.g. last week, month or year).

The term “**data collection period**” refers to the time period in which the information in the input data is collected. This may or may not be the same as the income/expenditure reference period.

The term “**base-year simulation**” refers to the case in which the year of policy rules matches the input data income reference period (e.g. in Tanzania, 2018 policy rules and Household Budget Survey 2017/2018 data). The purpose of base-year simulation is to provide information about the actual income distribution that is as accurate as possible. Thus, aggregate figures of simulated components should match as closely as possible the official statistics. However, this should be done without any calibrations of the sort that would distort the effect of simulated changes. The aim is to create the best starting point for simulating changes, not to reproduce base-year statistics as an end in itself.

The term “**target-year simulation**” refers to the case in which the year of policy rules does not match input data income reference period (e.g. 2018 policy rules and NIDS 2014 data). The purpose of target-year simulation is to provide accurate policy simulations for the policy year based on certain assumptions about growth in (market) incomes and/or expenditures in the input data (see Section 4) rather than deriving the actual income distribution for the policy year. Thus, aggregate figures of simulated and non-simulated components will not necessarily match official statistics.¹

The term “**baseline simulations**” refers to the best-possible combination between policy rules year and input data. See also Section 4.

Finally, EUROMOD software v2.0.5 and higher includes a new feature, the so-called “**Statistics Presenter**”. The tool lets the user pick an output dataset and calculates, based on this output dataset, the main indicators (government revenue and expenditure, inequality and poverty measures) of interest for the specified system (base system but also reform scenario). The tool also enables a comparison between base and reform scenarios. As of EUROMOD software v2.1.5 and higher the Statistics Presenter can also take into account indirect taxation (also see Section 12 in that regard). The Statistics Presenter requires a pre-defined set of variables and income lists from the output dataset. Refer to Section 2 on variables required in the input dataset, Section 7 for the required definition of poverty lines in the spine of the model, Section 10 for underlying income concepts, and Section 11 for required income lists.

¹ Nevertheless, when comparing against official figures one needs to check that the results from the simulation and uprating (see Section 4) move in a sensible direction and be able to explain any large variation in the model estimates. Furthermore, the Country Report discussion of the macro validation results should explain, where possible, the main reasons for diverging external and model estimates. See Section 14 for more on validation.

Desirable

An agreed two letter acronym is to be used (e.g. in tax unit and input/output file names etc)².
These are as follows:

BO – Bolivia
CO – Colombia
EC – Ecuador
ET – Ethiopia
GH – Ghana
MZ – Mozambique
PE – Peru
RW – Rwanda
TZ – Tanzania
UG – Uganda
VN – Viet Nam
ZM – Zambia
ZN – Zanzibar

² These are based on [ISO 3166 Alpha-2 country codes](#).

2 Input datasets

Essential

The input dataset must be kept in (tabulated) **text format** and the file named **CC_year_a#** where *CC* is country acronym, *year* refers to the data collection year, *a* is a letter standing for 'version' and *#* is the version number (1, 2, 3 etc.). For example, *tz_2018_a7.txt* refers to data for Tanzania (version 7) collected in 2018. The next revised version of the input dataset based on the first data source would be called *tz_2018_a8.txt*. If a second input dataset, based on a different data source, for 2018 was to be used, it would be called *tz_2018_b1.txt*.

All variables in the input data must be documented in detail in a Data Requirement Document (DRD). The DRD should be named to correspond to the input dataset file name, and the log page should be updated if a change is made. Such documentation needs to explain how the variables have been derived from the original source data and what they contain. See also information requirements in the DRD template. The DRD is critical for consistent documentation of the input datasets across the country models, such that any model can be understood and utilised by users outside of the country teams.

The variables used can be categorised as shown in Table 1, with more details offered below. For any variable that is required, either by EUROMOD software, Statistics Presenter, or specific country models, a variable with appropriate default values needs to be created when required information is not available (please refer to the country-specific DRDs and DRD template for default values).

Table 1. Variables used in SOUTHMOD models

Category	Variable type	Variables	Required
(1) Standard variables required by EUROMOD executable and queries	Identification variables (household, person, partner, mother, father)	<i>idhh, idperson, idpartner, idmother, idfather</i>	Always required
	Demographic variables (household head, age, gender, marital status, education, disability, household weight, country code)	<i>dhh, dag, dgn, dms, dec, deh, ddi, dwt, dct</i>	
	Labour market variables (economic status, occupation)	<i>les, loc</i>	
(2) Variables required by SOUTHMOD Statistics Presenter³	Standard equivalence scale (defined on model if possible)	<i>ses</i>	Always required
	Household expenditure	<i>xhh</i>	Always required for consumption based distributional statistics
	Imputed value of own produce	<i>xivot</i>	
	Base-year disposable income	<i>yds</i>	
	Some income variable(s) with positive values	<i>yem, yse or yag</i>	Always required for income based distributional statistics
(3) Variables required to model indirect taxes in SOUTHMOD	Variables capturing expenditure items	<i>x*</i>	Required for calculation of VAT
	Quantity variables	<i>q*</i>	Typically required for calculation of excise duties
	Price variables	<i>p*</i>	May be required, varies by country
	Variables for imputed indirect taxes	<i>tvaiv, texiv</i>	
(4) Variables that support model harmonization and cross-country analyses	Supplementary demographic variables (educational variables as per ISCED, citizenship)	<i>dec01, deh01, dcz</i>	May be required, depending on the purpose of modelling
	Supplementary labour market variables (industry as per different categorisations, formality, civil servant status)	<i>lindi, lindi01, lfo, lcs</i>	<i>lindi01</i> is required for modelling COVID shocks in 2020–2021, others depending on the purpose of modelling
	Additional equivalence scales (per capita and square root scales)	<i>ses01, ses02</i>	<i>ses01</i> is typically required for cross-country comparisons
	Original id variables from the underlying survey	<i>idorigperson, idorighh</i>	Not used on model but should be included
Other country-specific variables	Any variables required for the modelling of country-specific policies		

1. Standard variables required by the EUROMOD executable

Every dataset must include certain variables that are referenced by the executable. That is, EUROMOD checks if the variables are defined before reading the input data and will not run (or will issue a warning) if the variables are missing. Where no required information is available, a variable with appropriate default values needs to be created. These variables are:

i) identification variables

- *idhh*: household id number; numbered consecutively from 1 to *n*.
- *idperson*: individual id number; $idperson = idhh * 100 * i$, where *i* is the unique number of the individual in the household. The data must be sorted by *idhh* and *idperson* (ascending).
- *idpartner*: id number of spouse or partner if present in the household (cohabiting or married partner of the person in question); if not present set the variable to 0.
- *idmother* and *idfather*: individual id numbers of mother or father (biological or otherwise identified as female/male guardian) if present in the household; if not present set the variable to 0. As of EUROMOD version 3, *idparent* is no longer supported.⁴

ii) demographic variables (must be defined as non-monetary in the Administration of Variables tab on-model)

- *dhh*: household head; 1 for the head, 0 for all other household members.
- *dag*: age in years.
- *dgn*: gender; 0 for female, 1 for male.
- *dms*: marital status⁵;
 - 1: Single
 - 2: Married
 - 3: Separated
 - 4: Divorced
 - 5: Widowed
- *dec*: current level of education; as defined by EUROMOD;
 - 0: Not in education

³ In addition to the standard equivalence scale (*ses*), Statistics Presenter makes use of specific identification variables (*idhh*, *idperson*, *dwt*, *dag*, *dgn*, *dhh*); poverty lines (*spl*, *splpf*); and income lists (*ils_tax*, *ils_taxind*, *ils_sic*, *ils_bch*, *ils_bsa*, *ils_bsu*, *ils_bdi*, *ils_bun*, *ils_pen*, *ils_dispyx*, *ils_dispyx_pf*, *ils_con*, *ils_con_pf* – see Section 8 for details). The two lists *ils_dispyx** make use of income variables (generally at least *yem*) while lists *ils_con** make use of *xhh*, *xivot*, and *yds*. Statistics Presenter displays an error message if one of these four lists is left empty (or it makes use of unspecified income/consumption values) and is still used for the basis of calculating distributional statistics.

⁴ For African SOUTHMOD models, in particular, it is good practice to assign children with no parents (“loose children”) to a primary caregiver or household head if there is no information available on the primary caregiver. In many of the underlying survey datasets used with the African models, relationships within households cannot be readily identified and require imputations to assign children to adults. This requirement helps to ensure that such imputations have been carried out. In some models, loose children are not assigned to adults but dealt with when creating tax units.

⁵ Children should be included in this variable since some may marry before the age of 18.

- 1: Pre-primary education
- 2: Primary education
- 3: Lower secondary education
- 4: Upper-secondary education
- 5: Post-secondary education
- 6: Tertiary education
- deh: highest level of education achieved; as defined by EUROMOD;
 - 0: Not completed primary education (incl. those currently in primary education)
 - 1: Primary education
 - 2: Lower secondary education
 - 3: Upper secondary education
 - 4: Post-secondary education
 - 5: Tertiary education
- ddi: disability; 1 if the individual has a disability, 0 if the individual does not have a disability (the definition of disability is country specific). If in a country model levels of disability are needed, introduce a country-specific variable, named for example ddilv, for that purpose and describe its categories in the country report and the DRD.
- dct: country code; based on numeric [ISO 3166-1](#) country codes.⁶
- dwt: household weight; needed to provide population-level estimates. Observations with zero (or negative) household weights must be dropped.

iii) labour status variables

- les: economic status; as defined by EUROMOD; not required in SOUTHMOD but system required, set to missing (-1) if relevant information is not available. If set to missing the built-in query IsInEducation cannot be used in SOUTHMOD models as the query is based on variable dec > 0 and variable les = 6;
 - 0: Pre-school
 - 1: Farmer
 - 2: Employer or self-employed
 - 3: Employee
 - 4: Pensioner
 - 5: Unemployed
 - 6: Student
 - 7: Inactive
 - 8: Sick or disabled
 - 9: Other

⁶ Bolivia (68), Colombia (170), Ecuador (218), Ethiopia (231), Ghana (288), Mozambique (508), Peru (604), Rwanda (646), Mainland Tanzania and Zanzibar (834), Uganda (800), Viet Nam (704), Zambia (894).

- loc: occupation; as defined by EUROMOD; not required in SOUTHMOD but system required, set to missing (-1) unless a deliberate decision was made to define the variable in a particular country model. If set to missing the built-in query IsBlueColl cannot be used in SOUTHMOD models as the query is based on variable loc = 6 (skilled agriculture) or 7 (craft worker) or 8 (plant operator) or 9 (elementary occupation) or 0 (armed forces);
 - 0: Armed forces
 - 1: Senior officials and managers
 - 2: Professionals
 - 3: Technicians and associate professionals
 - 4: Clerks
 - 5: Service and sales workers
 - 6: Skilled agricultural workers
 - 7: Craft and trades workers
 - 8: Plant and machine operators
 - 9: Elementary occupations
 - -1: Not applicable

For additional (e.g. country-specific) definitions of loc and les, use loc01, loc02, les01, les02, etc., and similarly for other variables.

2. Variables required by Statistics Presenter

The Statistics Presenter tool requires additional variables to be defined in the input dataset and documented in the DRD:

- ses: equivalence scale as nationally defined for the household; value attributed to the household head, 0 for all other household members. This variable is preferably constructed 'on model' (in a policy within the model) for transparency reasons. Where this is not possible (due to lack of information about the equivalence scale), the variable should be added to the input dataset.
- xhh: household expenditure as used by national statistical offices for calculation of expenditure-based poverty. xhh might also include, for example, imputed values for self-produce (either estimated or as reported by respondents).
- xivot: imputed value of own produce (food and non-food) if available in data. The quality of data available will depend on the underlying dataset. This variable should be treated in a particular way in cross-country comparisons, as part of disposable income (also see Section 11).
- yds: base-year disposable income; derived from base-year simulation output and added to the input datasets. The variable is used to uprate consumption levels from the base year to the policy year (see Sections 10 and 11 for details).
- yem, yse or yag: employment, self-employment and agricultural income are part of income lists used by Statistics Presenter, which allow for calculating income based distributional statistics.

3. Variables for modelling indirect taxation

As of EUROMOD version 2.1.5 it is much easier to import expenditure, quantities and possibly price variables from the input data into the model, by checking the box 'Read expenditure related variables' under Country Tools – Databases (see Section 12 for more details on the treatment of indirect taxes and other additional information).

For the purpose of modelling indirect tax policies, the following variables should be created in the input dataset (based on the COICOP classification⁷) and documented in the DRD:

- variables capturing expenditure items start with the letter x followed by numbers relating to the COICOP classification (e.g. x01111 refers to cereals, x01112 to cereal flour); if an even deeper breakdown is required, two further digits can be added to the end of the variable name (e.g. local rice x0111104 and imported rice x0111105).
- quantity variables (typically required for calculation of excise duties) start with the letter q followed by the COICOP number of the respective item (e.g. q01111 for cereals).
- price variables (if needed) start with the letter p followed by the COICOP number of the respective item.
- variable names for imputed indirect taxes and variations thereof if necessary;
 - tvaiv - imputed VAT
 - texiv - excise duty

If an expenditure item from the input data does not conform to the deepest level of the COICOP classification but other items do fit the deeper levels, this item should be ascribed to a sub-group at a higher level of COICOP that fits the expenditure description more broadly.

For example, assume there is an expenditure item in the underlying dataset called "bread and cereals". This item cannot be captured on the deepest level of the COICOP classification such as rice (a cereal), as it is composed of a mix of the cereal and bread and bakery products COICOP categories (01.1.1.1 and 01.1.1.3 respectively). Therefore the item "bread and cereals" should be captured as a sub-group of the higher COICOP level capturing cereals and cereal products (01.1.1) more generally. The variable is thus named x011101 with the last two digits indicating the newly created sub-group.

4. Variables that support model harmonization or cross-country analyses:

For model harmonization and cross-country comparative analyses, some basic variables should be defined, documented in the DRD, and relevant country-specific circumstances explained in the country report. If a variable cannot be adequately defined due to limitations of the underlying dataset, it should be set to missing (-1) and mentioned in the DRD and country report.

⁷ Following the 2018 revision of the [COICOP](#).

- deh01: highest level of education achieved; based on [International Standard Classification of Education \(ISCED\) 2011](#);
 - 0: No schooling
 - 1: Early childhood/pre-primary education
 - 2: Primary education
 - 3: Lower secondary education and/or upper secondary education
 - 4: Post-secondary non-tertiary education and/or short-cycle tertiary education
 - 5: Bachelor or equivalent
 - 6: Master or equivalent
 - 7: Doctoral or equivalent
- dec01: current level of education; based on [International Standard Classification of Education \(ISCED\) 2011](#);
 - 0: Not in education
 - 1: Early childhood/pre-primary education
 - 2: Primary education
 - 3: Lower secondary education and/or upper secondary education
 - 4: Post-secondary non-tertiary education and/or short-cycle tertiary education
 - 5: Bachelor or equivalent
 - 6: Master or equivalent
 - 7: Doctoral or equivalent
- dcz: citizenship:
 - 1: In possession of country's citizenship
 - 2: Not in possession of country's citizenship but has citizenship from same continent
 - 3: Other
- lindi: detailed industry; as defined by EUROMOD;
 - 0: Not applicable
 - 1: Agriculture and fishing
 - 2: Mining, manufacturing and utilities
 - 3: Construction
 - 4: Wholesale and retail trade
 - 5: Hotels and restaurants
 - 6: Transport and communication
 - 7: Financial intermediation
 - 8: Real estate and business
 - 9: Public administration and defence
 - 10: Education
 - 11: Health and social work
 - 12: Other

- lindi01: detailed industry; variable denoting industry categories based on national definitions, required in current SOUTHMOD country models excluding Ecuador. The variable should be set to missing (-1) if the individual is not assigned an industry based on national definitions or he or she is not in the labour market. In other cases, the variable takes a positive value, denoting industry categories that differ across countries. The variable is used to apply economic shocks from COVID-19 'on-model' in 2020 using a definitional 'lma_cc' policy.⁸
- lfo: formality status; 1 if formal, 0 if informal. This refers to main job and should be based on the first occupation and usual economic status. If usual occupation is not available, use status of last week/reporting period. Add lfo01 for secondary job if available. If possible, follow the [ILO operational definition of informality](#), particularly Section B of Box 2 ("Criteria adopted for harmonized ILO estimates of informal employment") and Figures 3 and 4.
- lcs: civil servant (if possible follow [ISCO-08 classification](#)); 0 for 'No', 1 for 'Yes'.
- Two further equivalence scale variables should be included in the input dataset to ease cross-country comparisons:
 - ses01: per-capita equivalence scale, and
 - ses02: square root equivalence scale.
- idorigperson and idorighh: original id variables; should be retained so that the input dataset can be linked to the original data source.

Reference units

Data must be provided at the individual level (with people grouped into households).

All income variables must be provided with gross values, i.e. before deduction of employee and self-employed social insurance contributions and any taxes but excluding employer social insurance contributions. Where these are not available, they must be imputed.⁹

Where expenditure data is only available inclusive of VAT, the VAT policy will need to take this into account.

Income and expenditure data must be expressed in monthly terms, i.e. divided by 12 if originally recorded in annual terms, regardless of the actual number of months of receipt.¹⁰

Monetary variables need to be presented in the national currency.

⁸ For details, see Lastunen, J. (2022). On-model adjustment of incomes during COVID-19 in SOUTHMOD tax-benefit microsimulation models. WIDER Technical Note 4/2022. UNU-WIDER: Helsinki, Finland.

⁹ If both gross and net values are recorded in the data, such values must be checked to see if these are reliable (e.g. gross > net; ratio between gross and net according to income source, personal characteristics and tax-benefit rules).

¹⁰ Where the latter is known, this information needs to be retained in the input dataset for improving the simulation accuracy of monthly-based policies (e.g. social insurance contributions).

Sample adjustments

Observations with zero (or negative) household weight need to be dropped from the sample. If observations with positive weights have been dropped (e.g. because they had missing information on income), recalibration of weights could be considered.

Level of operation

EUROMOD operates only with personal level variables.

Any monetary variable at the household level in the original dataset (e.g., capital income, family allowances, social exclusion benefits, inter-household transfers, taxes and social contributions) must be assigned to one person in the household, usually the household head. Components of market income can be divided between several persons where it makes sense. When the choice of the person(s) is not obvious, the value should be assigned to the household head.

If capital and property income are provided at the household level, they must be assigned to the household head (the underlying assumption being that in the case of three or more generation households it would probably be the household head who would retain this kind of income).

Household level expenditure data should be assigned to the head of household.

Any monetary variable related to housing at the household level in the original dataset (housing allowances, imputed rent, housing cost) should be assigned to the head of household.

Any non-monetary variable at the household level in the original dataset should be assigned to all the persons in the household, for example region in which the household resides (drg).

Other imputations

Missing values are not allowed. For non-applicable (N/A), -1 or 0 can be used (see DRD).

Level of detail

All cash incomes available in the original dataset need to be retained, including information on taxes and benefits to be simulated in the model (where this is available).¹¹ Capital gains and other lump-sum incomes (e.g. lottery winnings, severance pay) must be clearly separated from other incomes (e.g. dividends, interests etc) if possible.¹²

Income data should be as detailed as possible. Only the same type of income can be aggregated (e.g. earnings from the main and secondary job but not two unemployment benefits where one is means-tested and another not), on the condition that more detailed information is not needed for the model and unlikely to be used. Incomes can be aggregated up to class 2 (see Section 3), e.g. yem.

¹¹ For validation purposes.

¹² In order to be able to exclude them from the standard concept of disposable income. However, for validation (and possibly other) purposes, it is useful to include them in the input dataset rather than excluding altogether.

Where both detailed and aggregated variables are included (e.g. yem and yem*), the exact relationship must be documented.

As mentioned above, original ID variables (renamed to idorigperson and idorighh) should be retained so that the input dataset can be linked to the original data source.

Expenditure data should have sufficient detail to enable the modelling of current VAT and excise duty policies while also allowing for policy reform (see Section 12 for more information).

Desirable

If population changes between the income reference period and policy year are such that they compromise the validation of the model, re-weighting could be considered as an option.¹³

Negative income values for all earnings (employment, self-employment and agricultural income) should be included without any adjustments. Negative income values should then be recoded as zero within the model, in the policy Neg_cc.

¹³ Note though that ideally this requires information on how original weights were constructed.

3 Variable naming convention

Essential

Variables must be named following a naming convention that consists of a list of acronyms joined together in a predetermined order to build the variable's name. For a list of variables and acronyms currently used in EUROMOD see the model (Administration Tools > Variables). The naming convention applies to variables in the input dataset and variables introduced on model.

There are two classes of acronyms ordered hierarchically:

Class 1: one character that identifies the type of variable (**a**sset, **l**abour market, **d**emographic, **s**ystem, **y** for (cash) market income, **e**xpenditure, **b**enefit, **p**ension, **t**axes and contributions, in-kind income); the **id**-variables are exceptions as they have a two-character Class 1 acronym.

Class 2: two-character acronyms specific for each variable type (assets, demographic, etc). Each Class 2 acronym has a unique meaning within each variable type (Class 1), but it could be that the same acronym means different things across variable types (e.g. **ag** stands for age among demographic variables and agriculture in assets, taxes, market income and benefits). The Class 2 acronyms are listed in ordered groups.

For example, employment income is named yem: **y** for market income + **em** for employment.

All variable names must always begin with a Class 1 acronym followed by at least one Class 2 acronym.

The order of acronyms in the variable name must follow the order of the groups in which these acronyms are included (to browse existing variables in EUROMOD, see *Administration Tools > Variables > Add Variables*). This prevents the same variable having different names because the acronyms are used in different orders. For example, a disability benefit for children must be named bchdi (b - benefit, ch - child, di - disability) instead of bdich.

It is recommended that acronyms from the same Class 2 group are not used together. If more than one acronym is used from the same group then these should be ordered alphabetically.

Intermediate Class 2 groups can be omitted, i.e. acronyms from a previous group are not compulsory.

It is not recommended to use more than five Class 2 acronyms together (this would occur in a name with more than 11 characters). Variable names typically include one to three Class 2 acronyms.

Each acronym should add relevant or useful information.

Before creating a new variable, it must be checked whether that name (or something similar) is already defined in EUROMOD. It is strongly recommended to use the existing variables whenever possible. For example, if there is already a variable that measures the time worked in months, avoid creating a new one that measures that in weeks or years.

When new categorical variables are created, the full list of categories, types or status must be documented in the DRD.

4 Uprating factors

Essential

Where income reference period and policy year do not match, monetary variables need to be uprated to the policy year. This is done by using relevant uprating factors in the model.

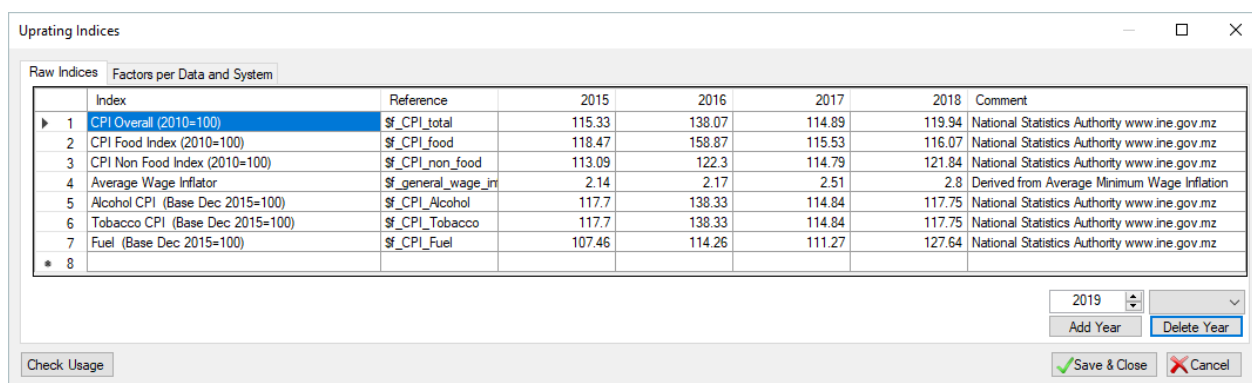
As a minimum, all monetary variables should be uprated using the CPI (or a factor of 1 if no change was observed).

It may also be possible to uprate income from employment/self-employment separately using another factor such as growth in average earnings.

The model constructs uprating factors for each input dataset on the basis of the raw data, taking the relevant year as the base period. Table 2 sets out an example of the raw data series for prices and growth in average earnings. These are added to the Uprating Indices tool accessed through Country Tools in the model.

The names of uprating factors referenced in the Uprate_cc policy (see below) start with the prefix \$f_.

Table 2: Raw data used for calculating uprating factors (example for Mozambique)



Index	Reference	2015	2016	2017	2018	Comment
1 CPI Overall (2010=100)	\$f_CPI_total	115.33	138.07	114.89	119.94	National Statistics Authority www.ine.gov.mz
2 CPI Food Index (2010=100)	\$f_CPI_food	118.47	158.87	115.53	116.07	National Statistics Authority www.ine.gov.mz
3 CPI Non Food Index (2010=100)	\$f_CPI_non_food	113.09	122.3	114.79	121.84	National Statistics Authority www.ine.gov.mz
4 Average Wage Inflator	\$f_general_wage_in	2.14	2.17	2.51	2.8	Derived from Average Minimum Wage Inflation
5 Alcohol CPI (Base Dec 2015=100)	\$f_CPI_Alcohol	117.7	138.33	114.84	117.75	National Statistics Authority www.ine.gov.mz
6 Tobacco CPI (Base Dec 2015=100)	\$f_CPI_Tobacco	117.7	138.33	114.84	117.75	National Statistics Authority www.ine.gov.mz
7 Fuel (Base Dec 2015=100)	\$f_CPI_Fuel	107.46	114.26	111.27	127.64	National Statistics Authority www.ine.gov.mz
* 8						

Desirable

If there is available information, separate uprating factors should be provided for food and non-food expenditure items.

5 System and database configuration

Essential

Monetary policy parameters should be in the national currency.

Output data must be in the national currency.

Any dataset or policy system (or part of it) which is not publicly available (e.g. any system under construction or system constructed for a specific project) must be defined as private.

6 Scope of policies

Essential

Policies are simulated as at 30 June (or, in the case of certain countries whose financial year begins at 1 July) in the corresponding policy year.

Where possible, and if they exist, policies to be simulated (at least partly) are:

- social insurance contributions: employer, employee, self-employed and credited contributions;
- income tax (excluding taxation of capital gains);
- property, wealth and other personal direct taxes;
- cash benefits (including unemployment benefits, family benefits, social assistance, education benefits, agricultural benefits such as farm input subsidies and direct subsidies more broadly speaking);
- in-kind benefits (if a suitable value of the benefit is available); and
- VAT and excise duties.

Policies are to be simulated in as much detail as possible given the underlying dataset.

Within-country regional differences are to be simulated as far as possible.

In-kind policies are to be simulated as long as a specific monetary value can be assigned to the benefit at the level of eligibility, such as the level of the individual or the household, for example school meals provided per school day for every child enrolled in school worth a specific value and food baskets with a clearly defined value distributed to the population.¹⁴ In-kind benefits must be clearly marked as such in the comments column belonging to the respective policy. Benefits are assumed to be cash benefits unless they are defined as in-kind benefits.

Direct subsidies are to be simulated if a clear value can be assigned at the level of eligibility, such as the level of the individual or the household. This must be clearly marked in the comments column belonging to the respective policy. Examples for direct subsidies are farm input subsidies or subsidies for utility costs, such as electricity and water. Indirect, price-based subsidies handed in first place to the producer of a household consumption good are currently not modelled.

Desirable

In general, policies must be implemented assuming full benefit take-up and no tax evasion. Depending on the country context it may be decided though to model non-payment of taxes and social security contributions by informal workers. In such cases this should be clearly indicated in the country report. However, whenever the macro validation exercise shows that results produced by the model substantially deviate from comparable external sources, non

¹⁴ Sufficient information on the policy and the input data must be provided for modelling either of the examples.

take-up and tax evasion could be modelled. The baseline, in such cases, will be based on the parameterisation accounting for non take-up and evasion. Both versions must be validated and documented. Where non take-up is modelled, random assignment is preferred over conditioning take-up on observed data receipt (i.e. households simulated to be eligible should be randomly selected *not* to take-up the benefit).

7 Policies

Essential

Policy simulations should begin with the following sequence: uprating factors (`uprate_cc`), constant definitions (`constdef_cc`) (if used – see Desirable section below), income list definitions (`ilsdef_cc`, `ildef_cc`, `ildef_stats_cc`, `ildef_exp_cc`), tax unit definitions (`tundef_cc`), recoding negative values of income (`neg_cc`), if used, and poverty line definitions (`spl_cc`).¹⁵

Policy simulations should end with the following two policies: individual-level standard output (`output_std_cc`) and household-level standard output (`output_std_hh_cc`).

All monetary values (including those on monthly basis) must have the period defined.

To store intermediate results in the baseline systems, define temporary variables for intermediate results (preferably, using a prefix `i_*` or `temp_*` in the variable name). The use of temporary variables with meaningful names (e.g. `i_benamt` for benefit amount) makes the model more transparent and readable.

Whenever joint taxation is applied, the simulated tax must be allocated proportionally to the taxable income/tax base between the members of the assessment unit in the relevant policy (where the income tax is simulated).

Each simulated benefit must be assigned to the most appropriate person within the assessment unit (often the head).¹⁶

Policies required by the Statistics Presenter

A poverty line policy – `spl_cc` – should be included. Most countries define two poverty lines, a higher and a lower line with the latter often referred to as the extreme or food poverty line. In the comments section, the name used for each poverty line must be clearly stated.

By default, the Statistics Presenter uses the benchmark poverty line of the country together with the national equivalence scale. This way Statistics Presenter tables reflect poverty and inequality results that are close to the commonly known levels in the country. The second poverty line can be selected by the user through an extension (see Section 8).

Furthermore, the Statistics Presenter requires each poverty line to be expressed in standard (`spl`) and in post-fiscal terms (`splpf`). Variable `splpf` is defined as the poverty line net off indirect taxes.

In order to calculate the distributional statistics, the Statistics Presenter also requires an equivalence scale variable called `ses` which should be calculated/defined on model in the equivalence scale policy `ses_cc`. By specifying the equivalence scale variable through a policy in the model, users can revise the definition and test a variety of equivalence scales.

¹⁵ Two important notes: First, uprating factors (implemented using `uprate_cc`) must be applied before any monetary variables are used. Second, the household head is defined through the first income list in the standard income list policy. This is relevant, for example, for the definition of the equivalence scales.

¹⁶ Assigning all benefits to one particular person allows the number of recipients to be counted more easily.

By default the nationally defined equivalence scale (often kilojoule/calorie-based) should be modelled, providing information on the specific weights in as much detail as possible on model. For ease of the user a per-capita scale (variable `ses01`) and a square root scale (variable `ses02`) should also be included in the input dataset but not included in the `ses_cc` policy. While the nationally defined scale aims to facilitate results that speak as closely as possible to the national context, the latter two scales are more suited for cross-country comparisons.

Desirable

A policy's name should ideally use the main output variable name of the policy (without `_s`) followed by `_cc` (e.g. `tscee_tz`). In the comments column, the policy's title must start with either DEF, SWITCH, TAX, BEN or SIC. The title must specify the name of the policy both in English and in native language in parentheses.

It is recommended that each independent instrument is modelled in a separate policy, e.g. every family benefit independent of others.

It is strongly recommended to store as many policy parameters as constants as possible and especially those repeatedly used throughout the model. Constants used in more than one policy must be defined in a separate constant definition policy (`ConstDef_cc`).

Where (short-term) benefits are adjusted with the number of months in receipt, this should be modelled as the last step in the relevant policy (if possible). Where social insurance contributions contain fixed amount elements, these should be adjusted with the number of months in work.

8 Switchable policies (extensions)

There are several 'policy extensions'. Extensions offer various additions to the default system for a policy year that users can select. More than one policy as well as functions from different policies can belong to a single extension. Furthermore, the same policy or function can belong to more than one extension. Extensions can be by default "switched on" (i.e. calculations are carried out) or "off". The default is defined in the "Set Switches" menu. Note that users can select whether to run the tax-benefit simulations with the extension being on or off in the run dialogue.

Adjustments, for example for limited benefit roll-out or tax non-compliance, must be done by adding switches (extensions), unless motivated differently in the country report. That way the policy system runs with these either switched on or off without requiring any further modifications and one is not required to have two separate systems and/or policies. Extensions included in a model can be found under Country Tools > Admin Country.

The standardised policy extensions depicted in Table 3 should be defined when needed:

Table 3: Standardised policy extensions

Extension Name (short name and long name)	Set to OFF	Set to ON	Baseline value (if exists)
BRO Benefit roll-out adjustment	Full benefit roll-out/delivery	Adjustment for incomplete roll-out or take-up (based on external data)	Country-specific
TCA Tax compliance adjustment	Full tax compliance	Adjustment for incomplete tax compliance (based on income tax or survey data, or relevant assumptions on payment by formality status)	Country-specific
FYA Full-year adjustment	Policies as of 30 th June or 1 st July	Annual policies (reflecting policy changes over the calendar year)	On
POV Choice of poverty line	Benchmark poverty line (generally nationally recognized)	Alternative poverty line (e.g. extreme or food poverty)	Off

Additional information on each extension is provided below:

- **Benefit roll-out adjustment (BRO):** In general, policies must be implemented assuming full benefit take-up. In some cases, results produced by the model may substantially deviate from comparable external sources, which suggests that adjustments may be required. When set 'on', this extension adjusts for incomplete benefit roll-out or take-up based on external data, such as information on actual tax receipts. Baseline applications vary by country, and country-specific choices should be discussed in the respective country reports.
- **Tax compliance adjustment (TCA):** In general, tax and social security contribution policies must be implemented assuming full compliance, with payments made by all liable taxpayers. As above, results produced by the model may at times deviate from comparable external sources, which suggests that adjustments may be required. Depending on the policy and target group (e.g. employed or self-employed workers), payments may need to be restricted only to individuals affiliated with the formal sector or otherwise adjusted to reflect actual payment amounts or the number of taxpayers from external data. When set 'on', this extension adjusts for imperfect tax compliance based on relevant assumptions or external information. Baseline applications vary by country, and country-specific choices should be discussed in the respective country reports.
- **Full-year adjustment (FYA):** This extension is applied specifically to the 2020 (and in some cases 2021) policy systems in standard models to adjust for the duration of COVID-related policies.¹⁷
- **Choice of poverty line (POV):** This extension enables users to choose the poverty line used when producing model output, namely the calculation of poverty rate and poverty gap.

¹⁷ See also [Gasior, K., Barnes, H., Joste, M., Lastunen, J., McLennan, D., Noble, M., Oliveira, R.C., Rattenhuber, P. and Wright, G. \(2021\). Full-year adjustment for modelling COVID-19 policies In SOUTHMOD tax-benefit microsimulation models. WIDER Technical Note 18/2021. Helsinki: UNU-WIDER.](#)

9 Tax units

Essential

The name of each tax unit must start with a prefix "tu_".

Tax units used in more than one policy must be defined in the tax unit policy (TUDef_cc).

For the purpose of marginal tax calculations, individual and household tax units must be defined (i.e. tu_individual_cc and tu_household_cc, respectively).

10 Income and expenditure concepts and relationship with income lists

A range of different income and expenditure concepts is used in the model, its income lists and subsequently Statistics Presenter. This section briefly defines different income concepts and how they relate to different income lists. The next section elaborates in detail on the components of the different income lists used in the model and by the Statistics Presenter.

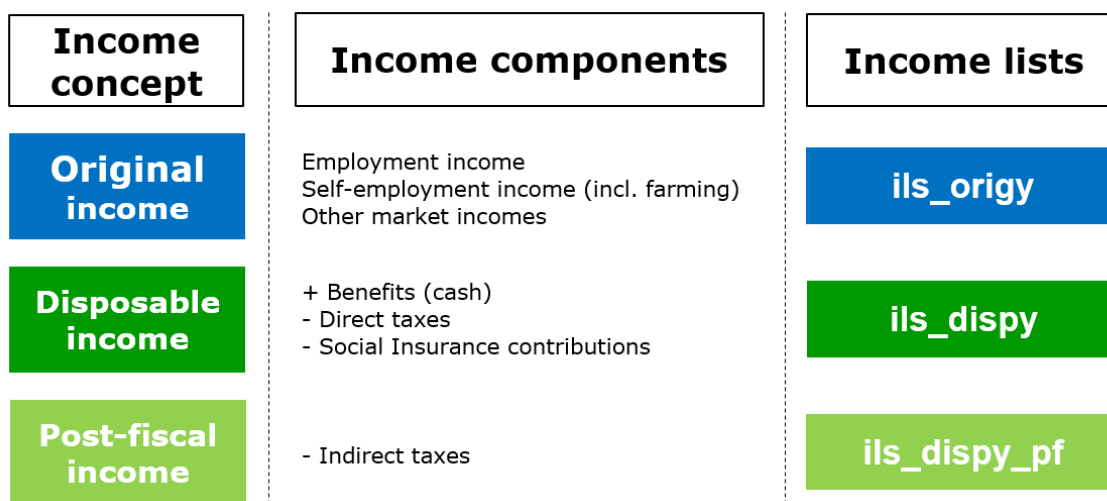
Figure 1 shows a summary of the key income concepts on the left hand side. First is **original income** (in dark blue), which is made up of employment income, self-employment income and other market incomes. In EUROMOD/SOUTHMOD terminology the income list for this concept is called `ils_origy`, and this represents the income before any benefits have been received and before any taxes have been paid.¹⁸

Disposable income (in dark green) represents the income *after* receipt of benefits and *after* payment of direct taxes and social insurance contributions. In EUROMOD/SOUTHMOD terminology the income list for this concept is called `ils_dispy`. An additional income list on disposable income accounting for in-kind benefits, `ils_dispyki`, is also included in the models.

Post-fiscal income (in light green) refers to disposable income after the deduction of indirect taxes. In SOUTHMOD terminology the income list for this concept is called `ils_dispy_pf`.

As own produce represents an important source of welfare in many developing countries, the income lists `ils_dispyx` and `ils_dispyx_pf` also include own produce and are by default used in the Statistics Presenter for income based poverty and inequality measures (also see Section 2 on the caveats when using own produce variables from the underlying data sources).

Figure 1: Overview of income/welfare concepts used for distributional analysis



Source: based on Gasior et al. (2018), page 6.¹⁹

¹⁸ In the Commitment to Equity (CEQ) fiscal incidence terminology, original income is referred to as 'Market Income'. 'Net Market Income' is defined as Market Income minus personal income and payroll taxes ; and 'Disposable Income' is defined as Net Market Income plus direct transfers.

¹⁹ [Gasior, K., Leventi, C., Noble, M., Wright, G. and Barnes, H. \(2018\). The distributional impact of tax and benefit systems in six African countries. WIDER Working Paper No. 2018/155. UNU-WIDER: Helsinki, Finland.](#)

Finally, as many developing countries use consumption to measure poverty and inequality, we introduce the concept of **simulated consumption** which is picked up by the Statistics Presenter.

The estimation of simulated consumption requires calculating the difference between (1) disposable income arising under the tax and benefit system in the base-year simulation, and (2) disposable income in the policy system of interest to the user. Both concepts include in-kind benefits. For ease of the user, disposable income defined as in (1) is provided as variable `yds` in the input dataset and is calculated during the first run of the model.²⁰ Disposable income as defined in (2) is simulated during the run of the model.

The consumption adjustments are carried out in the `xhhadj_cc` policy based on two assumptions. First, increases in disposable incomes result in the same increases in consumption. Second, decreases in disposable income result in decreases in consumption but account for consumption of own-account produced food. In these cases, a proportion of 25 per cent of the original consumption is assumed to be unaffected, based on Tschirley et al. (2015).^{21,22}

²⁰ The first run refers to a preliminary run of the model that generates simulated output variables that are used as input variables in the final underlying dataset. See also Section 2.

²¹ [Tschirley, D., T. Reardon, M. Dolislager, and J. Snyder \(2015\). The rise of a middle class in East and Southern Africa: Implications for food system transformation. *Journal of International Development*, 27\(5\): 628–646.](#)

²² See also Lastunen, J. (2022). On-model adjustment of incomes during COVID-19 in SOUTHMOD tax-benefit microsimulation models. WIDER Technical Note 4/2022. UNU-WIDER: Helsinki, Finland.

11 Income lists

Essential

Income lists should be organized in four separate income list policies:

- `ilsdef_cc` - for standard income lists;
- `ildef_cc` - for non-standard model specific income lists;
- `ildef_stats_cc` - for income lists only required for the Statistics Presenter; and
- `ildef_exp_cc` - for income lists collecting different expenditure items on the two-digit level of the Classification of Individual Consumption by Purpose (COICOP) classification.

All income lists used in more than one policy and/or defined as essential (see below) must be defined in the respective income list policy, others which are relevant only for a single policy can be defined in the corresponding policy. If an income list is part of the standard income list policy but also used by Statistics Presenter, it should sit with the standard income list policy.

The name of each income list must start with a prefix "il_" (for country-specific income lists) or "ils_" (for standard income lists and income lists required by the Statistics Presenter).

All income lists must have a description (next to where they are defined).

With regard to benefits, income lists should either include the benefit amount simulated in the model or the variable on benefit amounts collected directly from the data, but never both together. If both the benefit amount collected in the data and the simulated amount are available, preferably use the simulated amount (if simulated sufficiently well).

Make sure that no variable is double counted in income lists, e.g. a benefit component is included separately and at the same time as an aggregated variable or both data and simulated variable are included simultaneously.

Income lists should include detailed incomes rather than their aggregates (here referring to variables, not other income lists). For example, assuming there are two unemployment benefits and both need to be included, it is better to have each component separately rather than their aggregate.

Standard income list policy (ilsdef cc)

The standard income lists that should be included in the standard income list policy (`ilsdef_cc`) are listed in Table 4.²³ The colour coding in the leftmost column refers to colour coding that should be used for the respective income lists by default.

²³ Note that the income list for earnings (`ils_earn`) may contain several components that make up total employment income, such as main income, bonuses, commissions, overtime pay and work-related income benefits, as is the case in selected Latin American SOUTHMOD models. Furthermore, in addition to the three categories of social insurance contributions (SICs) listed in Table 4, an income list for other SICs (`ils_sicot`) can be used, as is the case for the Colombian model.

Table 4. Standard income list policy (*ilsdef_cc*)

	List	Contents	Notes
Yellow	ils_head	Household head – definition of the household head; if not defined, EUROMOD will by default define the household member with the highest income as the household head	Required for the model to run; must be included
Blue	ils_earns	Earnings – i.e. market income from employment (employment, self-employment and agricultural income)	Recode negative earnings to zero
	ils_origy	Original income – i.e. market income – has the following components (if available): <ul style="list-style-type: none"> • Earnings (+) • Income from capital, e.g. dividends and interests (+) • Income from occupational and private pensions (+) • Income from property (+) • Income received by children (+) • Regular inter-household cash transfer received (+) • Regular inter-household cash transfer paid (-) 	Required for the model to run; must be included Should exclude capital gains and other lump-sum incomes (e.g. lottery winnings)
Red	ils_tax	Direct taxes – composed of all direct taxes; identical to <i>ils_taxsim</i> if all taxes are simulated	Also used in the Statistics Presenter
	ils_sicee	Employee social insurance contributions (SICs) – also including compulsory private pension contributions	Include also SICs due on benefits and paid by the benefit recipients in <i>ils_sicee</i> , but in separate variables from employee contributions
	ils_sicer	Employer social insurance contributions (SICs) – composed of (employer) payroll taxes	
	ils_sicse	Self-employed social insurance contributions (SICs) – composed of self-employed social insurance contributions	
Green	ils_pen	Pension-related income – composed of all contribution-based pension benefits and non-contributory social pension benefits (simulated or not) but excluding old-age social assistance benefits	Also used in the Statistics Presenter
	ils_benmt	Means-tested cash benefits – excluding pension-related income (as in <i>ils_pen</i>)	Income- or proxy-means-tested
	ils_bennt	Non-means-tested cash benefits – excluding pension-related income (as in <i>ils_pen</i>)	Including severance pay
	ils_benki	In-kind benefits – composed of all in-kind benefits	Section 6 for definition
	ils_ben	Cash benefits – composed of all pension benefits, means-tested cash benefits, and non-means-tested cash benefits (<i>ils_pen</i> + <i>ils_benmt</i> + <i>ils_bennt</i>)	
Blue	ils_dispy	Disposable income not accounting for own produce: <ul style="list-style-type: none"> • Original income (<i>ils_origy</i>) (+) • Benefits (<i>ils_ben</i>) (+) • Direct taxes (<i>ils_tax</i>) (-) • Employee SICs (<i>ils_sicee</i>) (-) • Self-employed SICs (<i>ils_sicse</i>) (-) 	Include components if available
	ils_dispy_pf	Post-fiscal income not accounting for own produce: <ul style="list-style-type: none"> • Disposable income (<i>ils_dispy</i>) (+) • Indirect taxes (-) 	Include components if available
	ils_dispyki	Disposable income accounting for in-kind benefits: <ul style="list-style-type: none"> • Original income (<i>ils_origy</i>) (+) • Cash benefits (<i>ils_ben</i>) (+) • In-kind benefits (<i>ils_benki</i>) (+) • Direct taxes (<i>ils_tax</i>) (-) • Employee SICs (<i>ils_sicee</i>) (-) • Self-employed SICs (<i>ils_sicse</i>) (-) 	Include components if available

Statistics Presenter income list policy (ildef_stats_cc)

The income lists required (directly or indirectly) by the Statistics Presenter tool and collected in the Statistics Presenter income list policy (ildef_stats_cc) are listed in Table 5. The colour coding in the leftmost column refers to colour coding that should be used for the respective income lists by default.

Table 5. Statistics Presenter income list policy (ildef_stats_cc)

	List	Contents	Notes
Red	ils_taxind	Indirect taxes – composed of all indirect taxes	"Indirect taxes" in Statistics Presenter (SP)
	ils_sic	Social security contributions – composed of all social security contributions by employees, self-employed (if applicable) and employers	"SSC (employer, employee and self-employed)" in SP
Green	ils_bch	Child-related benefits – composed of all child-related benefits (including e.g. school feeding)	"Child benefits" in SP
	ils_bsa	Social-assistance related benefits – composed of all social assistance-related benefits	"Social assistance benefits" in SP
	ils_bsu	Orphan and widowhood-related benefits – composed of all orphan and widowhood-related benefits	"Orphan/widow benefits" in SP
	ils_bdi	Disability-related benefits – composed of all disability-related benefits	"Disabled benefits" in SP
	ils_bun	Unemployment-related benefits – composed of all unemployment-related benefits	"Unemployment benefits" in SP
	ils_bag	Agricultural benefits – composed of all agricultural benefits	"Agricultural benefits" in SP
Blue	ils_dispyx	Disposable income accounting for own produce: <ul style="list-style-type: none"> • Disposable income (<i>ils_dispy</i>) (+) • Imputed value of own produce (food and non-food, variable <i>xivot</i>, see above) (+) 	Include components if available
	ils_dispyx_pf	Post-fiscal income accounting for own produce: <ul style="list-style-type: none"> • Disposable income accounting for own produce (<i>ils_dispyx</i>) (+) • Indirect taxes (-) 	Include components if available
	ils_con	Simulated consumption – consisting only of variable <i>xhh_s</i> , which is defined in a separate policy as follows (see Section 10 for details): <ul style="list-style-type: none"> • Consumption (<i>xhh</i>) (+) • Disposable income of policy system in focus, also including in-kind benefits (<i>ils_dispyki</i>) (+) • Disposable income in the base year, again including in-kind benefits (<i>yds</i>) (-) 	Variable <i>yds</i> captures disposable income, including in-kind benefits, in the base year and is uprated automatically through the uprate function
	ils_con_pf	Post-fiscal simulated consumption: <ul style="list-style-type: none"> • Simulated consumption (<i>ils_con</i>) (+) • Indirect taxes as simulated (<i>ils_taxind</i>) (-) 	

The first set of income lists required are listed first and coloured red (taxes and SICs) or green (benefits). Note that a benefit might fit in more than one of the income lists for benefits defined above for the Statistics Presenter. However, each benefit should only be attributed to one of the benefit income lists in order to avoid double-counting. For example, an orphan benefit is also a child benefit, in which case, choose the list `ils_bsu` as it matches more closely the benefit type. The income lists `ils_pen` + `ils_bch` + `ils_bsa` + `ils_bsu` + `ils_bdi` + `ils_bun` + `ils_bag` need to add up to `ils_ben` + `ils_benki`.

The second set of income lists required includes two different types of disposable income (`ils_dispyx` and `ils_dispyx_pf`) on top of the disposable income lists already defined above in the standard income list policy (`ils_dispy`, `ils_dispy_pf` and `ils_dispyki`).

These lists (disposable income and post-fiscal disposable income accounting for own produce, `ils_dispyx` and `ils_dispyx_pf`) are used by Statistics Presenter by default. However, users can easily set the variable `xivot` to "n/a" if they prefer to use disposable income and post-fiscal disposable income without accounting for own produce. Similarly, users can add `ils_benki` to these income lists to also account for in-kind benefits, akin to list `ils_dispyki` (see Table 4).

Income lists `ils_dispyx` and `ils_dispyx_pf` are mixing income and expenditure concepts. The reason is that in many developing countries own produce plays an important role and household income as captured in `ils_dispy` evaluates to zero for many households. Nevertheless there might be great differences between households with zero incomes in terms of own produce and hence welfare. See Section 2 on caveats when using variable `xivot`, particularly in a cross-country comparative manner.

The third set of income lists (`ils_con` and `ils_con_pf`) is required for countries that typically measure poverty based on consumption. The concept of simulated consumption is used to estimate the impact of tax-benefit policies on consumption. The underlying assumption is that all changes in disposable income (including changes in in-kind benefits) from the base year to the year in focus lead to changes in consumption by the same amount.

Income lists for expenditure items (`ildef_exp_cc`)

In the income list policy for household expenditure items (`ildef_exp_cc`), 13 income lists following the two-digit level of the COICOP classification are required (see more in Section 12 on the treatment of indirect taxes):

- `ils_coicop01` for items belonging to COICOP 01;
- `ils_coicop02` for items belonging to COICOP 02;
- etc.

Finally, there should be an income list collecting standard rated items subject to VAT called `ils_vat_std` which should sit directly within the VAT policy. In case of a two-tier VAT system, "`ils_vat01`" and "`ils_vat02`" should collect the respective expenditure items subject to the two different tax rates (and further lists need to be added in case of more than two tax rates).

Desirable

Non-standard income lists which could be included, depending on what policies exist, include taxable income for income tax purposes and the sum of incomes relevant for means-testing of benefits.

12 Indirect taxes

Essential²⁴

Expenditure variables and quantities are part of the general input dataset that contains the demographic, labour market and income variables. The expenditure variables and quantities are read in by the EUROMOD software by checking the box 'Read expenditure related variables' under Country Tools – Databases.

The various expenditure items are introduced to the model according to the COICOP classification (see Section 2 for more information). 13 income lists collect expenditure items relating to consumption expenditure at the household level (see Section 11 for more information). The goods and services subject to the current VAT policy are in turn collected in the income list `ils_vat_std` within the VAT policy for countries with one standard VAT rate. In case of a two-tier VAT system, "`ils_vat01`" and "`ils_vat02`" collect the respective expenditure items subject to the two different tax rates (and further lists need to be added in case of more than two tax rates). Income lists relevant for other indirect tax policy/policies are collected in the corresponding indirect tax policy/policies.

For the simulation of indirect taxes two or three types of variables (depending on the country's indirect tax policies) are prepared in the input data (see Section 2 for how variables should be named in the input dataset):

- expenditure data net of indirect taxes;
- quantities/units of goods in case of excise duties calculated per unit;
- values of indirect taxes that were removed during the data preparation stage, are brought as imputed values into the model.

Variable names created in the model (and variants thereof):

- `tva_s` for simulated VAT
- `tex_s` for simulated excise tax

Some items are subject to VAT and excise duties simultaneously. The correct order of application of the two taxes must be considered when stripping off indirect taxes and calculating the imputed VAT and excise duty.

Expenditure variables are automatically uprated by the `uprate` policy using the default factor unless otherwise specified. If variables need uprating by uprating factors other than the default factor, this must be undertaken within the `uprate` policy. Quantity variables are not, of course, uprated.

The excise duty policy should be undertaken at the individual level rather than the household level if some or all of the excise duties are based on quantities (i.e. if some or all of the excise duties are not *ad valorem*). This is because parameters that use non-monetary variables can

²⁴ Also see Adu-Ababio, K., Pirttilä, J., Rattenhuber, P. and Vanheukelom T. (2019). Integration of indirect taxation to GHAMOD. WIDER Technical Note 2021/12. UNU-WIDER: Helsinki, Finland.

only be applied at the individual level, not household level (see 'Parameter values and the assessment unit' page in EUROMOD help files).

Naturally, in the base year the imputed values of indirect taxes (calculated during the data preparation stage) must be the same as the values of indirect taxes calculated by the model. It is therefore important to compare tvaiv with tva_s and texiv with tex_s; they should be identical.

13 Output

Essential

Results must be outputted at the individual level by default.

Output (at the individual level) needs to include:

- all income variables from data which are not simulated in the model (including fringe benefits);
- all simulated income variables;
- all socio-demographic variables;
- all income lists;
- (desirable: identifiers for all tax units).

Make sure that output policies do not have any variable (or income list) twice or any temporary/intermediate variables. Naming intermediate variables starting with `i_name` and using them for debugging and checking purposes is good practice. This simply requires specifying `i_name` variables in the output policy.

14 Validation

Essential

It is important to undertake as many of the following validation steps as possible.

Micro validation

Check eligibilities and the amounts of taxes and benefits simulated by the model (case-by-case validation for a selection of particular households);

Compare simulated values against data recorded values in the same survey on case-by-case basis;

Check descriptive statistics for outcome variables (min, max, mean);

Check that results for some basic indicators (e.g. average tax rate) make sense for all observations in the sample;

Macro validation

Compare the sum of each income component and the number of recipients (both original incomes and tax-benefit instruments) with external statistics; simulated values can be also compared against data recorded values in the same survey. For simulated instruments, this should answer the question how close the model estimates are to the external figures given the following factors:

- a) quality of external statistics;
- b) survey quality, i.e. how representative market/non-simulated incomes and population structure are, measurement errors etc;
- c) quality of data imputations (e.g. net-to-gross imputations, income splitting, replacing missing values);
- d) simulation quality, i.e. accuracy of tax-benefit rules; and
- e) key modelling choices and assumptions (e.g. adjustments for benefit non-take up and tax evasion, the 30 June/1 July rule, etc).

In general, the focus should be on relative differences, e.g. how one instrument compares to another, whether the bias has an expected sign and whether trends over years are in the expected direction. Calculate and compare inequality measures (Gini, S80/S20 ratio) and poverty measures with external statistics. This should address the question of how different the model estimates are due to differences in the underlying methodology. The following sources of bias are justified (but need to be acknowledged):

- sample adjustments;
- differences in the concept of disposable income;
- data imputations, i.e. (c) above; and
- differences between data and simulated values, that is (d) and (e) from above combined.