EUROMOD annual meeting
New data process

2 1-22 September 2020

Content

- EUROMOD input data preparation after transition
- From the data sources to the EMSD
- From the EMSD to the EUROMOD input data
- Support documents
- Practical arrangements

EUROMOD input data preparation

(after transition)

NSIs

Creation of 2 datasets

1) EU-SILC

2) SILC

Data transmission

EC Regulation (2019/2242) and Delegated Regulation (EU) 2020/25

Practical Arrangement (PA)

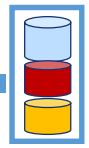
eurostat

Provision of yearly parameters

Creation of 4 datasets

Creation of Euromod country-specific var (type N)

EMSD transmitted according a PA + EC Regulation 557/2013 to cover UDB component



1) UDB var

2) SILC var

3) EM var derived by Estat (type E)

1) PDB

2) UDB

Provision of Euromod country-specific var

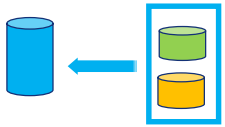
3) EMSD

Interaction to create the EM final input data

Joint Research Centre JRC

Creation of the 4th dataset

4) Euromod final input data



AT CZ EE EL IE IT LT LV LU NL PL SI SK UK

Data flow tested with CZ ES and LV NTs

ZOOM - EMSD file structure

1) UDB data

2) SILC national var. produced by NSIs

3) EM var. derived by Estat

idhh	idperson	db090	rx040	hy050g	py010g	hy020_nsilc	V6_G	Reg	dag_e	dms_e	les_e	yem_e	amrrm_e	afc_e	xhc_e
7255400	725540001	1526	0,611111	1392	20349,01	2500	565	6	55	2	5	1695,8	4	0	435,875
7255400	725540002	1526	0,611111	1392	46320	2500	100	6	50	2	3	3860	4	0	435,875
7255400	725540003	1526	0,611111	1392	6721,11	2500	0	6	19	1	3	560,09	4	0	0
7255400	725540004	1526	0,611111	1392	0	2500	45	6	16	1	6	0	4	0	0
7255500	725550001	1141	99	0	37515,6	670	700	4	60	2	3	3126,3	5	3333,33	130,495
7255500	725550002	1141	0	0	0	670	356	4	58	2	9	0	5	3333,33	130,495
7255600	725560001	1407	99	0	0	3500	0	1	67	2	4	0	10	644,444	465,04
7255600	725560002	1407	0,285714	0	38550,89	1500	0	1	56	2	3	3212,6	10	644,444	0
7255700	725570001	1177	1	0	36293,21	300	658	3	58	4	3	3024,4	3	1288,89	229

Fixed content
Var. defined in EC Regulations

Content defined in PA
Var. vary among MS

Content defined with NTs & JRC
Var. vary among MS

From the data sources to the EMSD

- Step 1 – *Gathering the various data*

  UDB and PDB in Estat

  National data (+ ID maps): practical arrangement with NSIs

  Yearly parameters & relevant statistical source(s) of info (input from NTs)

- Step 2 – *EMSD preparation*

From the data sources to the EMSD

Ex. Yearly parameters needed from the Latvian NT

EM variable	Parameters needed (SILC 2019 survey year)	Source of information
afc	interest rate and source of average interest rates ('avir' variable)	
dec, decde, deh, dehde	school age thresholds by level of education	
dew	Age when country's levels of education are attained	
dey	Number of years spent in the country's levels of education	
les	1. retirement age for both men and women 2. school age thresholds	
yiytx yiynt	A tax rate on income from capital	

From the EMSD to the EM input file

- Step 1 - *Categorisation of EUROMOD variables*

  Estat's method to classify EM var. between type E,E+p and N

  → complexity, stability of the code between years

  → need country-specific knowledge

  → limit back and forth between Estat and NTs to prepare the EM input data

  Fuzziness for some var: type "E" or "N" ?

  Comparison of Estat's results with NT's sumstats: source of discrepancy?

  → NT's feedback on the classification, rationales for changes

From the EMSD to the EM input file

- Step 1 - *Categorisation of EUROMOD variables*

  Learnings from the pilots (ES,LV – CZ work in progress)

CZ (103 - 10) ES (103 - 30) LV (115 - 18) before NT's assessment

→ agreement with almost all Estat classification

→ LV: 4 var. changed from type N to E/E+p (labour and income)

→ ES: changes from type E to N: Idfather,Idmother,Idpartner – inconsistencies which need manual corrections

changes from E to E+p and vice versa (education level, labour)

From the EMSD to the EM input file

- Step 1 - *Categorisation of EUROMOD variables*

  Concerns / questions from NTs

  $?_1$ Do I have to use EM var. computed by Estat?
  → not mandatory but loss information from the PDB var.

  $?_2$ How can we identify a EM var. type
  → in EMSD data: variable name suffix '_e'
  → in DRD partially filled by Estat (DRD_EMSD_CC_Yxx.xls)

From the EMSD to the EM input file

A	B	C	D
Labour market information		Compres Expand	
Type	Variable	Label	Notes: derivation from original data, and comments
E	lcs	LABOUR MARKET : Civil Servant 1 yes 0 no	gen lcs = . replace lcs=1 if temp_occup==1 | temp_occup==2 | temp_occup==3 | temp_occup==11 | temp_occup==23 | temp_occup==34 | temp_occup==54 replace lcs=0 if lcs==. where: temp_occup=pl051
N	liwwh	LABOUR MARKET : In work : Work history (length of time in months)	liwwh = pl200 * 12 No of observations with missing or invalid work history (liwwh):1544. No of observations with null or n/a labour information and positive related income: 1003. Missings values are imputed using age when began first regular job
E	loc	LABOUR MARKET : Occupation (ISCO 1-Digit) 0 Armed forces 1 Senior officials and managers 2 Professionals 3 Technicians and associate professionals 4 Clerks 5 Service and sales workers	loc = 0 if pl051 == 1 | pl051 == 2 | pl051 == 3 loc = 1 if pl051 == 11 | pl051 == 12 | pl051 == 13 | pl051 == 14 loc = 2 if pl051 == 21 | pl051 == 22 | pl051 == 23 | pl051 == 24 | pl051 == 25 | pl051 == 26 loc = 3 if pl051 == 31 | pl051 == 32 | pl051 == 33 | pl051 == 34 | pl051 == 35 loc = 4 if pl051 == 41 | pl051 == 42 | pl051 == 43 | pl051 == 44 loc = 5 if pl051 == 51 | pl051 == 52 | pl051 == 53 | pl051 == 54 loc = 6 if pl051 == 61 | pl051 == 62 | pl051 == 63 loc = 7 if pl051 == 71 | pl051 == 72 | pl051 == 73 | pl051 == 74 | pl051 == 75

From the EMSD to the EM input file

A	B	C	D	E	F	G	H	I
Variable	**Obs**	**Weight**	**Mean**	**Std. Dev.**	**Min**	**Max**		**Type**
aca	34.738,00	45.755.145,30	1,27	0,64	1,00	3,00		E
aco	34.738,00	45.755.145,30	1,38	0,75	1,00	3,00		E
afc	11.441,00	14.444.113,90	60.660,12	262.238,90	258,62	9.279.655,00		E+p
amrrm	34.738,00	45.755.145,30	4,92	1,30	1,00	10,00		E
amrtn	34.738,00	45.755.145,30	2,14	1,20	1,00	6,00		E
ate	34.738,00	45.755.145,30	1,03	0,23	1,00	3,00		E
bch								N
bch00								N
bchdi								N
bchot								N
bed	459,00	595.564,03	149,85	146,82	2,50	1.170,00		E
bfa	338,00	426.354,23	197,67	216,57	3,33	2.200,00		E
bhl	531,00	664.139,95	304,03	348,39	1,48	2.979,05		E
bhl00	473,00	600.950,20	277,31	338,87	1,48	2.979,05		E
bhlot	58,00	63.189,75	558,11	337,82	13,33	2.100,00		E
bho	326,00	455.084,12	135,33	136,86	4,00	750,00		E
bma								N
bsa	558,00	759.666,76	338,56	284,68	5,00	1.777,27		E
bun								N

DRD_EMSD (sumstats sheet): cell shade, column I

From the EMSD to the EM input file

?$_3$ Will type E/E+p var. be derived with the same formulas and imputation methods than the ones used in previous years by NTs?

→ yes unless new recommendations on the default imputation method from JRC

Type	Variable	Label	Notes: derivation from original data, and comments
E	amrrm	ASSETS : Main Residence : Number of Rooms	gen amrrm = hh030 **Imputation:** * Missing values are replaced using the average ratio between household size and number of rooms gen temp_ratio = hh030 / hx040 sum temp_ratio replace amrrm = round(r(mean) * hx040) if amrrm == .
E	amrtn	ASSETS : Main Residence : Tenure 1 Owned on mortgage 2 Owned outright 3 Rented 4 Reduced Rented 5 Social Rented 6 Free 7 Other	gen amrtn=hh021 recode amrtn (1 = 2) (2 = 1) (5 = 6) **Imputation:** * Missing values are replaced with aco (min) mode values sort idhh egen temp_mode_amrtn = mode(amrtn) if idhh != idhh[_n-1], minmode qui sum temp_mode_amrtn gen average_amrtn=r(mean) replace amrtn = average_amrtn if amrtn== .

From the EMSD to the EM input file

?$_4$ Shall we assess the EM var. categorization every year?

→ no, one-off task, minor changes between years

→ new EM year, quick review of E-type and N-type lists

?$_5$ What should I do to resort to E-type var. to compute the other input var. ?

→ step 2 of “how to go from the EMSD to the EM input file” !

From the EMSD to the EM input file

- Step 2 - *Deriving EUROMOD var. defined as type N*

EM N-type var. mainly in labour and income blocks (exceptions ∃ for some MS – IT, EE)

Personal information	EM var. created by Eurostat
Labour market	EM var. created by Eurostat
	Countric-specific EM var. generated by the NT
Income, Benefits and Taxes information	EM var. created by Eurostat
	Countric-specific EM var. generated by the NT
Value of assets	EM var. created by Eurostat
Expenditures	EM var. created by Eurostat
Eurostat	EM var. created by Eurostat

From the EMSD to the EM input file

- Step 2 - *Deriving N-type EUROMOD var. (or following the current process)*

  Single data source : EMSD

  → No need to merge ≠ datasets, no ID maps issues

  → Additional info from PDB in E-type var.

E-type var. included in EMSD

→ Var name = EM name_e (ex. *dms_e*)

→ Generated with same formulas and imputation methods as the NTs (or following JRC recommendations)

→ Data handling already performed (agreed at the kick-off meeting): no missing values and/or inconsistencies

From the EMSD to the EM input file

- Step 2 - *Deriving N-type EUROMOD var.*

  Learnings from the pilots (LV, ES – CZ work in progress)

  → New process to generate the EM input file from the EMSD:

  easier

  allows the NT to focus its efforts on var. that need more in-depth analysis to derive them

  → Assessment of the var. categorization (E/E+p/N) run in parallel of N-type var. derivation

From the EMSD to the EM input file

- Step 2 - *Deriving N-type EUROMOD var.*

  Learnings from the pilots (LV, ES – CZ work in progress)

  → Methodology to update the do-files

  1. Identify do-files where N-type var. (and temp var. used for N-type) generated → do-files selection
  2. Use of E-type var.

  → delete or comment code lines to produce them

  → replace E-type var. name in the do-files (gen dms=dms_e)

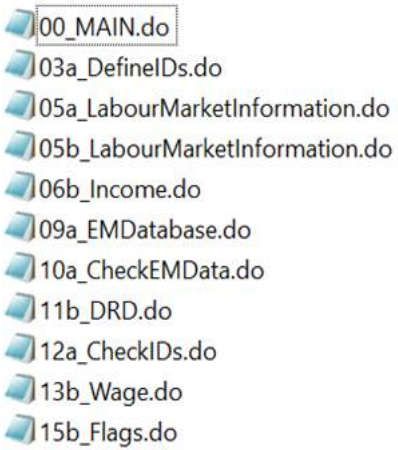
  3. Update the selected do-files to prepare the N-type var. & EM input data

From the EMSD to the EM input file

→ Impacts of the use of the EMSD

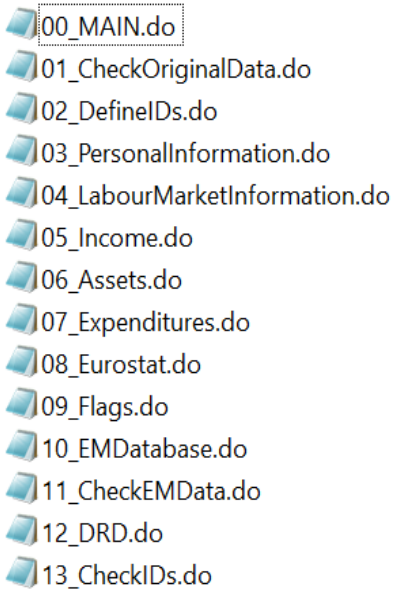
1. Simplify the code and the checks: important issues of E-type and EU-SILC var. double checked (Estat and NT)
2. Reduction on the number of scripts

LV (115 - 18)



11 vs 27

ES (103 - 30)



14 vs 28

From the EMSD to the EM input file

- Step 2 - *Deriving N-type EUROMOD var.*

  Concerns / questions from NTs

  $?_1$ Will Estat provide the do-files used to generate E-type var. ?
  → no, Estat works with SAS but ∃ code lines in DRD

  $?_2$ Can NTs change the structure of the progammes to take into account the use of E-type var. ?
  → yes, delete/comment do-files creating E-type var., group do-files to generate N-type var.

From the EMSD to the EM input file

?$_3$ Can NTs double check important issues leading to warnings or error messages (ex.drop children born after the income ref period) ?

→ yes, EMSD here to make NT's work easier

?$_4$ What to do if I want to produce all EM input var. ?

→ take the EMSD as the new data source and run your do-files

Support documents

1. Usual documents (updated)

   Derivation of EUROMOD Input data from EU-SILC

   EUROMOD Modelling Conventions

   Metadata on the disaggregated benefits

   Do-files templates (no changes to take into account E-type var.)

   Excel file with new changes (log_dofiles_drd_2020-01-28.xls)

   DRD template + DRD partially filled by Estat

Support documents

2. DRD partially filled with Estat's var.

   DRD template as a basis, same layout

   Sheet "Guide" with explanations

   EUROMOD var. identified (E, E+p, N) – column, cell coloured

   E-Type var. : code lines to generate them, summary statistics

   N-Type var. : code lines as in the template, no sumstats

   → NTs complete

   option 1: the DRD partially filled with information on N-type var. (code used, summary statistics)

   option 2: "traditional" DRD if no use of E-type var.

Support documents

3. Metadata to explain the EMSD

   Background information on the categorisation of EM var.

   New data preparation process

   Data handlings agreed with NT at the kick-off meeting

   Tables:

   - yearly parameters needed
   - national data included in EMSD
   - PDB var. used inside Estat
   - all E-type EUROMOD var. derived with PDB var. (or nat. data)

- Learnings from the pilots (LV – CZ,ES work in progress)

→ documentation sufficient to prepare N-type var. and EM input data

Practical arrangements

- Kick-off meeting to launch Y12 in Nov-Dec 2020 (NTs-JRC-Estat)

  Yearly parameter for EM Y_{12} (available in CR)

  Data handling practices

  → Correct (or not) the warnings and inconsistencies identified regarding IDs, age and sex

  → Correct (or not) the ID numbers when personal ID does not correspond to the Household ID (the cases of split-off households - 01c_FixSplitIDs.do file)

  → Set a upper limit to “dew” to correct for any date in the future

  → ...

Practical arrangements

- Kick-off meeting to launch Y12 in Nov-Dec 2020 (NTs-JRC-Estat)

  Presentation of EM var. categorisation but assessment when preparing EM input file

  Careful look at E-type var. when discrepancy between NT and Estat summary statistics and no persuasive explanations

  Need to review imputations of type E/E+p var. before EMSD release ? (from ES pilot)

- Platform: S-CIRCABC
- EUROMOD data preparation timeline

EUROMOD data preparation timeline

Oct – Nov

UDB available inside Eurostat

Eurostat

Nov-December

- Kick-off meetings EM Y_{12}
- EMSD preparation

NTs-JRC-Estat

Jan - Feb

EMSD and support documents preparation

Eurostat-JRC

Feb-March

- Launch of EM Y_{12}
- EMSD available to NTs + documentation

NTs-JRC-Estat

April

EM input files for the FE

NTs-JRC (Estat)

August

EUROMOD internal release

NTs-JRC-Estat

December-January

EUROMOD public release

JRC-Estat

Practical arrangements

- Features for EUROMOD Y12 (2019 SILC data)

  Contents of EMSD

  → no extra national data, start with 2020 SILC data (Y13)

  → but E-type var. based on PDB var.

  Situation per country

BE BG CY DE DK ES FI FR HR HU MT PT RO SE

AT CZ EE IT SK

(C) EMSD with 2 components

- UDB var.
- Type-E var.

EMSD complemented with national data in Y13

(A) EMSD fully implemented

EM input data based on
- UDB + full national SILC
- national SILC only (IT SK)

Condition: GA between JRC-Estat-NSI signed before Dec 2020

(B) EMSD partially implemented

EM input data based on UDB + subset of national SILC

EMSD complemented with additional national data in Y13

(D) No EMSD in Y12

EM var. not yet classified (E/E+p/N)
Countries not replicated
Some MS may move to (A)

NL LV

IE SI EL LT LU PL UK

Before leaving

Contact: ESTAT-EUROMOD@ec.europa.eu
or Henri - Veronica - Albane

- Thank you for your help since 2018
  - to understand the EM input data preparation
  - to get further national var.
  - to test the new data preparation process
- Time for questions !