



7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

D. Data processing: Effective IT strategies for data processing (including LFS files for Eurostat) and data storage

D1 – Eurostat data processing: current challenges and possible future improvements

Frank Espelage – Eurostat



EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social Statistics
Unit F-3: Labour Market



7th Workshop on LFS methodology, Madrid, 10-11 May 2012

Eurostat's LFS data processing: current challenges and possible future improvements

Frank Espelage (Eurostat)

3 main objectives

- Short overview of Eurostat's LFS data validation and editing procedures in order to foster common understanding for Eurostat requests to countries
- Some examples of problems Eurostat faces regarding data validation
- Future developments in the area of data validation which might solve some issues mentioned before

Background

- Eurostat needs harmonised data for compilation of EU figures and comparisons across countries and over time
- LFS regulations define data to be collected, including formats for datasets, code lists for variables and routing. Also principles for the collection of important variables etc.
- Validation as joint task of NSIs and Eurostat: countries collect, validate nationally, transmit – Eurostat validates, compiles, disseminates. Certain validation steps are (at least theoretically) performed both by countries and Eurostat, others can only be done at Eurostat via comparisons across countries
- Several requests/initiatives to improve this situation:
 - o TF Quality LFS: share Eurostat's validation programs (SAS) for better understanding of requirements and harmonisation of validation
 - o Joint ESS strategy: more efficient and integrated production methods for statistics (horizontally and vertically)

Current situation

- For a given reference year, Eurostat receives up to 4 different kinds of datasets per country: quarterly, yearly, ad hoc module, household samples
- Different data(sub)sets are transmitted either together in one file or in separate files
- Standard processing order at Eurostat:
 - o Quarterly/yearly data
 - o Ad hoc module data

- Household data (for countries sampling individuals with special household subset)

How does Eurostat process transmitted data¹?

- Reception via eDamis – validation of quarterly data starts normally immediately after receipt of the dataset(s)
- File format OK? One line per record, data in correct columns etc.
- Valid codes in line with definitions in the regulation or common code lists like ISCO, NACE, ISCED, NUTS, country code lists?
- Routing respected, i.e. filter conditions in the regulation correctly implemented?
- Few checks on household composition, (too) few “soft” checks (e.g. age vs year of highest education attainment)
- Calculation of aggregated main results². Comparison with previous quarters (calculation of differences and visual check).
- Main results plus possible further info sent to NSI. Request for either confirmation or correction and retransmission of data

Processing can be divided in HARD checks and additional plausibility checks which could trigger further analysis of data. Hard checks include: file format, validity of codes, routing. As regards those criteria defined by regulations, there should basically be no errors!

Reality: still data editing needed to get valid codes / consistency across variables / valid routing, i.e. Eurostat CHANGES transmitted data in line with (partly self-defined) general rules, agreement with countries and/or based on previous experiences/consultations. In consequence, datasets used nationally and at Eurostat might differ.

Taking the trade-off between timeliness of dissemination and completely error-free data into account, Eurostat informs countries about such changes and asks for remedy of invalidities and inconsistencies at least for future transmissions of the same dataset, but accepts partly incorrect datasets for the time being if only few records show serious problems or proceedings regarding similar errors were already discussed before.

Examples:

- ISCO, NACE etc – attempt to find a valid higher level
- AGE and other values like YEARESID, YSTARTWK, HATYEAR etc should show a logical structure
- Variables not allowing the coding as blank (either transmitted blank or necessary recoding after routing checks): these are often the most important variables (identifiers like HHNUM, HHSEQNUM, demographic background, labour status etc). Some of the variables needed for the ILOSTAT calculation might be missing, so Eurostat defines them in a way neither creating artificial employment nor unemployment (trade-off between this unsatisfactory solution and dropping records plus possible reweighting)

¹ Cp for this: *Eurostat control of LFS microdata country transmissions* and *LFS – SAS control routines for NSIs*

² Main result calculations should be reproducible by using either the shared SAS validation programs or the descriptions of the calculations on the specific main results sheets explaining it. General information on Eurostat’s derived variables can also be found in the EU LFS User Guide.

- AGE calculation: difficult situation when last reference week of a reference year ends in the following calendar year. Datasets do not contain AGE, but REFYEAR, YEARBIR and DATEBIR, and DATEBIR is defined differently across countries, either relative to REFYEAR or CALYEAR.

Even if Eurostat normally informs NSIs about changes and asks for corrections/cleaning either before possible re-transmissions/revisions of the same data or for future datasets, it nevertheless receives data with similar or identical problems again.

Relations between different subsamples

Yearly data:

- Normally combined sample of the 4 quarterly datasets unless subsampling of structural variables
- In case of subsampling (“wave approach”): Eurostat implements “all or nothing” approach, i.e. one variable subsampled → all yearly variables will be analysed based on subsample and using yearly weights only. Other information is suppressed. Main reason for this approach is the already high complexity to manage all data peculiarities at Eurostat. What is more, many users do have serious problems already now to understand the distinction between the different datasets.
- Main result comparison: average of 4 quarters vs. wave approach (consistency of results for subgroups). Experience: correct weighting of the subsample not that easy, especially if further conditions/samples have to be taken into account (AHM, household).

Household data:

- Normally part of yearly data
- For countries sampling individuals: option to provide special household data for subset of individuals, using also special weights. Household information should then be combinable with all other yearly variables.
- In the latter case: main results comparison (rules however not that clear..)

AHM data:

- Generally processed AFTER (quarterly and yearly) core data.
- AHM variables should be combinable with ALL other variables, i.e. in case of subsampling for yearly and/or household variables the yearly/household samples should cover the AHM sample (not always the case yet..)
- Data transmission by countries in several different formats³
- At Eurostat, creation of AHM datasets for validation technically done via merge of selected identifiers⁴ + AHM variables with already validated core data (-> preference for this AHM transmission format!). Hard checks are performed on these datasets then.
- Detailed results distributed to countries.

³ The 3 main transmission formats are described in *LFS – SAS control routines for NSIs*

⁴ The identifiers used are HHNUM HHSEQNUM SEX YEARBIR REFYEAR COUNTRY (even if the first two should by definition already be unique and identical across different quarters/datasets for a given respondent)

Problem: sometimes difficult to decide whether filter condition is fulfilled or not. Examples:
AHM 2009 – simple text filter extremely difficult to transfer into code (TRANSACT).
AHM 2010 – sample of individuals vs. filter on existence of children...

Revisions

Eurostat receives lots of revisions⁵, triggered for instance by Eurostat requests for corrections to countries, detection of inconsistencies/errors by NSIs themselves, implementation of new classifications (NUTS), reweighting due to new census results, later transmission of variables filled with administrative data (education, INCDECIL) etc.

- If only few variables (with simple filters or little impact on other variables) are affected, often easier to get transmission of identifiers + specific variables than whole files. Advantages: allows "batch" processing, no concerns about possible changes in other variables
- Identifiers should be unique and identical across different quarters/datasets for a given respondent – allows comparison by record to detect changes in processed files
- Eurostat often detects more changes than announced/intended, hence preference for e.g. AHM transmission as identifiers + AHM variables. Future standard for AHM transmission?
- Revisions do not mean bad previous data, but could also be sign of high NSI interest in clean data!
- Bundling of revisions to make life easier?

Possible future improvements

Vertical integration

- Several initiatives aiming at validating data the earlier, the better -> all validation which can be performed at NSI level should be done there
- Upstream validation (via eDamis or separate GSAST tool) in preparation at Eurostat
- Checks for this upstream validation to be defined (LAMAS)
- Data will only be considered as transmitted if it arrived in Eurostat AND has passed the upstream validation before -> compliance monitoring!
- Time horizon?

Advantages:

- Eurostat and NSIs forced to agree on what are essential quality requirements
- Use of the same data (less or no further editing at Eurostat)
- Repetition of basic hard checks (after upload) not necessary anymore -> more time to develop and implement additional criteria and more sophisticated soft checks at Eurostat
- Can be improved and extended over time

⁵ In theory, Eurostat should receive some 130 quarterly files per year (33 countries x 4 quarters). In practice, Eurostat processes more than 500 quarterly datasets per year. Data for some countries and quarters have already been transmitted more than 10 times.