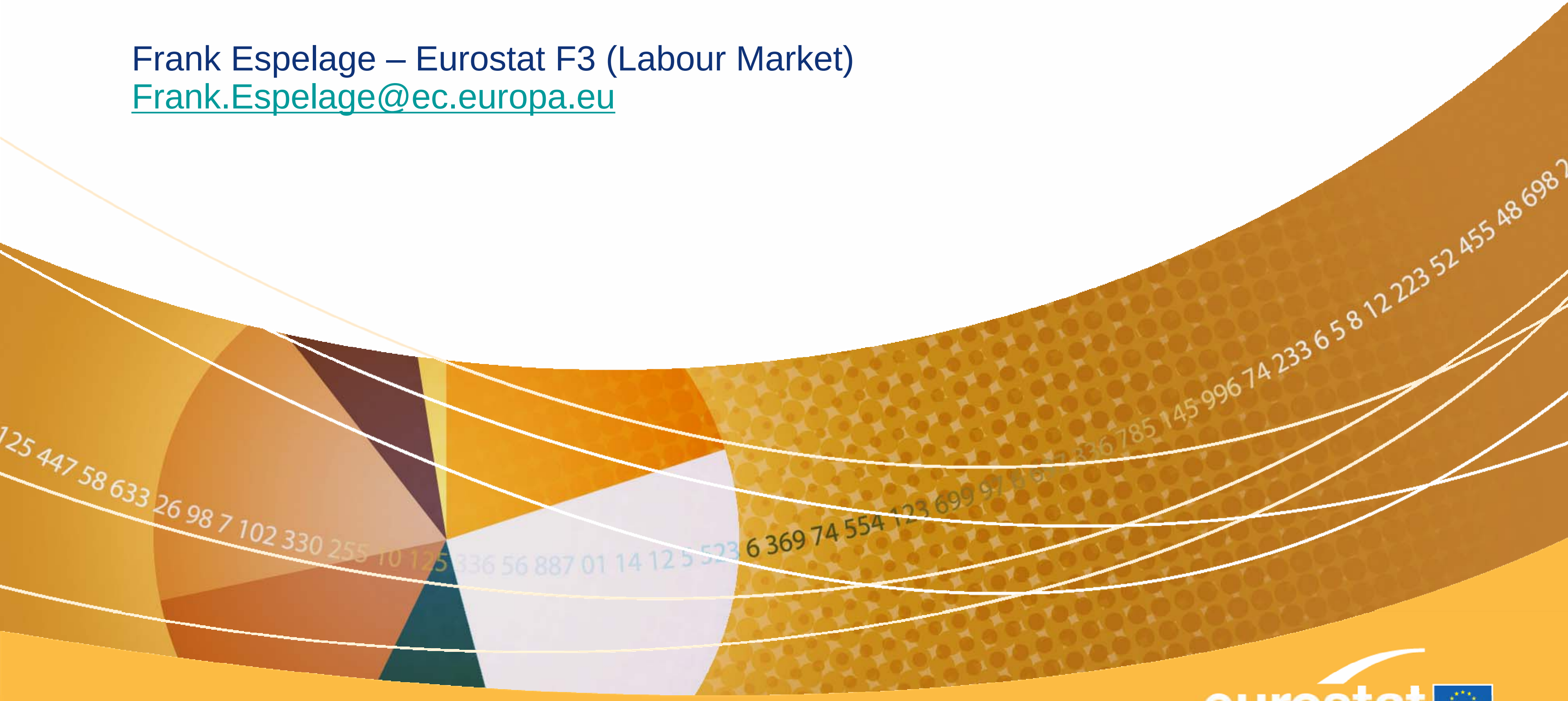
Eurostat's LFS data processing: current challenges and possible future improvements

Frank Espelage – Eurostat F3 (Labour Market)
Frank.Espelage@ec.europa.eu



Background

- Need for harmonised data at European level

- Validation as common responsibility of NSIs and Eurostat

- High importance assigned to validation:
  - joint ESS strategy for more efficient and integrated production methods for statistics, both horizontally across different domains and vertically across data providers and Eurostat
  - VIP Validation inside Eurostat, common procedures and tools across NSIs and Eurostat

Current situation – validation procedure

- Eurostat receives quarterly, yearly, ad hoc module, household datasets

- Processing according to data needs – quarterly for unemployment releases, yearly for structural analyses, ad hoc module / household data only after finalisation of other datasets

- Validation of file formats, validity of codes, routing, checks of household data, soft checks. Comparison of main results with previous data. Confirmation of NSIs sought.

- SAS programs available to NSIs for recent reference years.

Current situation – validation checks

- HARD checks – often defined by regulations – should not result in ANY errors, but

- Eurostat often detects invalid codes, inconsistencies across variables, invalid routing etc., hence

- Eurostat EDITS and CHANGES transmitted LFS data
  - no delay in dissemination of unemployment results or other headline indicators
  - in parallel, Eurostat informs countries about errors detected and asks for future rectification

- Other surveys do not accept incorrect transmissions, e.g. EU-SILC

Examples – simple errors

- Nationally valid, but not EU-LFS conform values for occupation, economic activity etc.
- AGE and other date/time variables like YEARESID are filled with values which do not show a logical structure (AGE << YEARESID)
- Recent EU level changes not implemented yet, 2012 example: EL instead of GR for COUNTRY, NATIONAL, COUNTRYB, COUNTRYW etc.

Examples – invalid blanks

Variables not allowing code *blank* delivered as *blank* or requiring a recoding from the transmitted *not applicable* to a valid code.

- Still the case for instance for labour status variables like WSTATOR, SEEKWORK, METHODS, AVAILABLE, but also others
- In such cases, Eurostat recodes WSTATOR to 5, SEEKWORK to 3, METHODS to 0, AVAILBLE to 2 and informs the country
- Reasoning for that recoding: do not create artificial effects, neither employment nor unemployment
- Alternative: deletion of records - would theoretically require reweighting by NSIs

Examples – difficulties to calculate correct AGE

AGE calculation for reference year ending the following calendar year:

Eurostat does not get AGE, but uses a fixed REFYEAR and given auxiliary variables YEARBIR and DATEBIR to calculate

DATEBIR='1' -> AGE=REFYEAR-YEARBIR

DATEBIR='2' -> AGE=REFYEAR-YEARBIR-1

with DATEBIR = 1 (2) if the person's birthday falls between 1 January and the end (after the end) of the reference week

Examples – difficulties to calculate correct AGE

- Most countries deliver data which allows Eurostat using its standard formula, but around 10 calculate DATEBIR relative to the calendar year and not the fixed reference year
- Creates problems if the reference year and the calendar year differ, i.e. when the last reference week of a year ends in the following calendar year (2004, 2009, 2010, 2011…)
- AGE calculated at Eurostat could be different from original age, leading to
  - wrong population distribution across age classes
  - wrong routing etc.
- Current "solution": adapt auxiliary variable YEARBIR (but see the consequences below)

Relations between different subsamples

Yearly data:

In case of subsampling of (some) structural variables – "all or nothing"

Household data:

For reference persons in special household datasets all quarterly and yearly variables should exist

AHM data:

Should be combinable with ALL other variables (yearly and household samples must cover the AHM sample)

At Eurostat: AHM variables and identifiers merged with validated core data -> strong preference for that transmission format !

Identifiers

- Identifiers (theoretical): HHNUM HHSEQNUM
- Should be unique even across datasets, but not yet always the case: often more variables needed to identify a respondent
- Identifiers (used): HHNUM HHSEQNUM SEX YEARBIR REFYEAR REFWEEK COUNTRY

- YEARBIR correction for AGE calculation causes possible problems here
- Stable identifiers allow better monitoring, for instance a comparison new / old by record in case of revisions

Revisions

- Transmission of complete datasets or subsets? Often easier to get identifiers + revised variables, in particular
  - if only few variables are affected and
  - a long time series has to be revised

  Examples: NUTS 2010, weight revisions after availability of census results
- Eurostat often detects more/less/other changes than announced
  - important to clarify with NSIs: why? Also needed to inform users correctly
- Should there be a policy of less but bigger revisions?
  - less work, but possibly
  - more differences between national and Eurostat figures

The (near) future - plans

- Validate data as early as possible in the production chain, i.e. all what can be done at NSI level should be done there

- Upstream validation in preparation at Eurostat
  - Checks for this upstream validation to be defined through LAMAS WG; will for sure contain basic regulation requirements
  - Future data transmissions will require clean data, otherwise the file will not pass the upstream validation and not arrive in Eurostat - compliance relevance!
  - High ESS priority: should be implemented the next 2-3 years

The (near) future - consequences

- Eurostat and NSIs forced to agree on quality requirements regarding data validation
- Improvement and extension over time possible, i.e. start with basic regulation requirements and add additional checks later
- In the long run, use of identical data at NSIs and Eurostat

Thank you for your attention