IT-LFS Microdata file under new EU regulation 1700/2019

Antonella Iorio, Claudio Falorni, Alessandro Martini, Silvia Montecolle, Federica Pintaldi

iorio@istat, falorni@istat.it, alessandro.martini@istat.it, montecol@istat.it, pintaldi@istat.it

Topic: Microdata production

1. **background**

According to the European regulation, micro validated data of the Labour Force Survey must be sent on both quarterly and annual basis according to the Eurostat explanatory notes. The deadlines for data transmission follow the survey calendar, so NSIs have to provide quarterly data within 8 weeks from the end of the survey period and by the end of March the data referring to the previous year.

Istat also guarantees, in compliance with current legislation, the access and release of validated microdata deriving from its own surveys for statistical-scientific and research purposes and LFS is historically one of the surveys with higher requests for access to elementary data.

Access to microdata for external users can take place in two ways: the release of a validated microdata file or, for users who need special processing for specific research needs, Istat makes the Laboratory for Analysis of ELEmentary Data (Laboratory Adele) available. This laboratory, located at the Istat headquarters and its regional offices, is a Research Data Center (RDC), is a safe place where researchers and scholars from universities, institutes or research bodies can directly process the data and carry out statistical analyzes on microdata from the Institute's surveys.

All the microdata available to external users, including those of the Labor Force survey, are archived centrally in the Armida system (MIcroDAti ARchivio) which is used to respond to requests from the bodies of the National Statistical System (Sistan) and is accessible at the Adele laboratory.

The researcher must present a research project and obtain permission to access the data from Istat's President, he must also sign a contract requiring him to protect statistical confidentiality.

Once the researcher has been given permission to process the requested data, the outcome is reviewed by an Istat researcher for confidentiality protection and, if possible, released.

The protection of data confidentiality at the Adele Laboratory takes place from three points of view: physical, since the workstations dedicated to the service are located within the Institute's offices and access to them is supervised and allowed only to authorized users. Furthermore, the workstations do not allow the entry or withdrawal of data and are isolated from the Internet network; legal, as the researcher undertakes in writing, making himself liable to report to the competent judicial authority, to respect the rules on statistical confidentiality and personal data protection; of statistical confidentiality, due to the control to which the results of the researcher's elaborations are subjected before their eventual release.

Microdata release responds to specific regulatory aspects that distinguish the different types of files that Istat disseminates, which respond to different information needs of several types of ursers.

The directive "Criteria and methods for the communication of personal data within the National Statistical System" of the Comstat of 20 April 2004 (Directive n. 9/Comstat), provides that a body or statistical office belonging to the Sistan can request another subject of the System personal data already acquired for statistical purposes. On the basis of this directive, Istat is required to communicate personal data collected through its surveys to the Sistan entities that request them in order to conduct their studies included in the National Statistical Program or for their institutional purposes.

SISTAN files are therefore not subjected to statistical procedures aimed at reducing the risk of identifying respondents, the information content is completely similar to the information processed by ISTAT and all aspects relating to the protection of personal data and guarantee of statistical confidentiality are under the responsibility of the organization that has accessed the data.

The elementary data created to satisfy scientific research needs are called MFR and have been regularly produced for the FdL survey since 2010. These are data files, with no direct identifying elements, to which control methods have been applied for the protection of confidentiality. Access to files can only be requested for the implementation of a specific research project by researchers belonging to organizations recognized as a research body recognized by Comstat, the Statistical Information Steering and Coordination Committee (Comstat), the governing body of the National Statistical System or by Eurostat.

2. Guidelines for the production of microdata files

When the data for the first quarter of 2021 were released, with the entry into force of regulation 1700/2019, the identification risk assessment was reviewed, following the most recent Institute guidelines.

A specific methodology is adopted for the preparation of the files for dissemination relating to the continuous survey on the Labour Force, in order to limit the risk of identification of the respondents.

The intrusion scenarios considered are:

• identification through external archives, or through the connection with data released by other public sources;

• spontaneous identification, i.e. resulting from a priori knowledge of the user which could allow the correct attribution of the data released to the units of the surveyed population.

The variables involved in the protection process are those that can allow the association between the information and the respondents, namely:

• uniquely identify the statistical units of detection/analysis (such as, for example, address and tax code);

• they allow to limit the population to which the respondents belong and, alone or in combination with others, can lead to the re-identification of one or more records.

While the former are deleted from the file, the latter are treated statistically by reducing their information content.

Therefore, initially all direct identifiers are removed, such as Name and Surname, etc., and the so-called work/control variables, i.e. relating to data collection methods (Time type of the interview, Name of the interviewer, etc.)

We move on to the identification of the key variables, i.e. those which, even for a single method, have at least one of the following characteristics:

rarity of the values in the population under study,

visibility of the character, or some of its modalities, by an observer,

traceability in external archives.

In the labour force survey, some of the main key variables to keep under control are: the municipality, or province, of residence, work or study, the state of birth or citizenship, marital status or profession.

To evaluate the identification risk, the sample frequencies of the combinations of each key variable by age group are first analysed, and then the combinations of two or more key variables.

Given the two parameters k e p, respectively:

$k_{MFR} \in \{2,3\}, p \in [0,0.1]$ for the MFR file e $k_{Micro} > k_{MFR}, p \in [0,0.01]$ for the public user file, the following rule must be satisfied:

$$\frac{n.\,individuals\ in\ distinct\ groups\ belonging\ to\ cells\ with\ frequency < k}{n.\,individuals\ in\ the\ data\ set} < p$$

When the previous condition is not satisfied, it is necessary to intervene using a specific statistical technique to protect the privacy of the respondents.

One possibility is represented by the global recoding in which the modalities of the key variables are merged taking care that the resulting classes are both compliant with those adopted in the official publications and maintain coherence with similar classifications adopted in the same survey context.

It is desirable that the new codings are maintained over time both for reasons of confidentiality and to allow the comparability of data collected on subsequent occasions. Global recoding, being by definition related to all records (even those that would not present critical issues in terms of confidentiality protection), can lead to a relevant loss of information detail.

Otherwise it is better to resort to local suppression: in correspondence with only the records that violate the rule, the mode of a key variable is set to missing.

3. The production of microdata files for IT-LFS

In the light of the results of the analysis described, the global suppression and/or recoding interventions were defined for the production of the MFR files starting from the first quarter of 2021.

It should be noted that any effort has been made to maintain the structure and information content of the MFR files produced until 2020, considering that users of the microdata file for scientific research purposes are still required to take the confidentiality protection measures according to their commitments signed with the data communication request.

Several variables, detected according to the survey questionnaire, have been made unavailable in the file. Among them, it is significant to mention the direct identification codes, the survey control variables and the municipal territorial references. Individual and family identifiers were generated through pseudo-random numbers to allow for individual and family analysis.

Global recodings concerned key variables such as age, marital status, family type and profession. For these variables, more aggregated classifications have been defined, capable of reconciling the need to safeguard the information detail and limiting the risk of identifying respondents.

The disclosure risk assessment also regarded the variable relating to gross salary, the estimation methodology for which is still being studied, but the preliminary results made it possible to update the recoding criteria.

For the preparation of the mIcro.STAT file relating to the continuous labor force survey, starting from the first quarter of 2021, an appropriate methodology was adopted that is consistent with the Institute's recent guidelines in this context. Methods applied now refer exclusively on local suppressions and global recoding of the data in order to improve consistency with the estimates disseminated by the survey, since the final weights are not recalculated. In the previous version of the public user file sampling were used to protect individuals with an high disclosure risk, this meant that consistency with the survey data was guaranteed only with respect to the variables included in the additional calibration. This was a pertinent problem for a

group of users not allowed to access the MFR file, since for instance figures for Profession or Sector of activity derived by PUF file were not coherent with official ones.

The direct identifiers and the territorial references at the sub-regional level have been eliminated, while for the sector of economic activity and the profession only variables recoded into classes and large groups have been released.

The global recoding, on the other hand, concerned age, spread across classes, the region of residence, with the unification of Valle d'Aosta and Piemonte, educational qualifications and family typology.

Similarly, some quantitative variables have been recoded into classes, sometimes wider than the corresponding variables released in the research file, such as for example the hours worked, detailed information on dates and durations, number of family members, etc. Further data protection interventions concern local suppression for some records through the insertion of missing values in correspondence with one or more variables, while for others, relating to individuals not belonging to the reference population of the survey, we proceed with the elimination from the data set of the entire individual record.

All the elementary data release are accompanied by exhaustive documentation that describes both the information content through the adaptation of the questionnaire to the specific version of the file and the technical and methodological aspects useful for carrying out the processing through a statistical package.

4. Back reviewed microdata for 2018-2020

It was also possible to provide users files of microdata prior to 2021. For the years 2018 to 2020, questions were introduced in the survey questionnaires that allowed a link between the old definition of employment and the new one introduced by regulation 1700/2019. This allowed for the backward reconstruction of some information that was made available to researchers allowed to access to SISTAN microdata files a reduced number of variables consistent with those released from 2021 onward.

Reconstruction of data prior to the introduction of the regulation was done at the macro level for the main indicators, such as employment, unemployment, and inactivity rates, for some socio-demographic characteristics such as gender, geographic areas, and age groups. Reconstruction at the micro level from 2018, allowed for some specific indicators that ISTAT produces (e.g., the non-participation rate or the share of involuntary part time).

Information reconstruction was done for those who resulted employed according to the new definition but were not employed under the old one (about 0.04% of the total individuals for year 2018). Since most of the cases are women who have been on parental leave for more than three months, some information was reconstructed in a deterministic way, also retrieving it from the section on “Previous work experience”. Others, on the other hand, were imputed by probabilistic methods when the small number of cases allowed. In the opposite situation (about 0.09% of the total individuals for year 2018), employed according to the old definition and unemployed according to the new, the information was placed as missing.

For many variables it was not possible to reconstruct the information at the micro level. This happened when the changes introduced in the new questionnaire were such as not to allow a comparison with the information collected according to the old regulation.

5. Conclusions

With the production of the microdata files for the IT-LFS survey according to the new EU Regulation 1700/2019, an attempt was made to improve the quality of the released data. On the one hand we tried to ensure continuity and coherence with the previous series. A significant improvement is the better coherence of the public user file with the official estimates applying mainly global recoding to protect individuals with an high disclosure risk though this leads to a certain loss of information detail.

References

ISTAT, 2006, La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione. Istat, *Metodi e Norme*, No. XX.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. e de Wolf, P.-P. (2012). Statistical Disclosure Control. Wiley.

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica (Metodi e norme, n. 20, 2004), http://www.istat.it/dati/catalogo/20040706_00/.

Willenborg, L. e de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 111, New York: Springer-Verlag