EUROPEAN COMMISSION
EUROSTAT

Directorate A: Resources
Unit A-5: Methodology; Innovation in official statistics

ESTAT.A.5/AB

ACCESS TO CONFIDENTIAL DATA FOR SCIENTIFIC PURPOSES

(SCIENTIFIC USE FILES)

GUIDELINES FOR PUBLICATION

Commission européenne, 2920 Luxembourg, LUXEMBOURG - Tel. +352 43011

Estat-microdata-access@ec.europa.eu

CONTENTS

2. INTRODUCTION

These guidelines explain how confidential data for scientific purposes are to be used by researchers[1]. The researchers must read the guidelines before they use confidential data and respect the rules for publication laid down below. In accordance with the paragraph 5 of the "terms of use of confidential data" signed by the duly designated representative of the research entity, the researchers are bound for publication by the following guidelines.

Researchers must ensure that all research published or otherwise disseminated does not contain information that allows individual statistical units (persons, households, enterprises, etc.) to be identified.

In all reports, including both published and unpublished papers, researchers must ensure they have abided by the strict application of the guidelines for publication attached to the confidential data for scientific purposes.

(Paragraph 5 of the terms of use of confidential data)

3. WHAT IS CONFIDENTIAL DATA?

Data are confidential if the respondents can be identified. Simply removing name and address details from the microdata files does not prevent the identification of the survey respondents.

Specific or unique characteristics of the survey respondent may lead to their recognition. The scientific-use files delivered to researchers are especially prepared to make the identification of survey respondents more difficult. This is done by:

- reducing the level of detail of the data;
- modifying certain values;
- and/or suppressing risky records or variables.

4. WHY IT IS IMPORTANT TO PROTECT CONFIDENTIALITY (IDENTITY) OF THE RESPONDENTS

We collect data from individual respondents to produce statistics. Researchers may be granted access to files containing information on individual respondents (access to microdata) to conduct statistical analysis for scientific purposes.

[1] Acknowledgements: "Responsible Access to ABS Confidentialised Unit Record Files (CURFs); Training Manual"; Australian Bureau of Statistics; March 2005.

Confidential data for scientific purposes contain information provided by individuals or organisations. Each record in the microdata files represents information provided by the respondents.

Researchers granted access to confidential data are only permitted to use the data to conduct statistical analysis for scientific purposes.

Researchers are prohibited from identifying individuals or organisations represented in the files.

Disclosure of individual information constitutes a breach of the law, but also a breach of the trust the respondents place in the statistical system. Such a breach could harm the reputation of that statistical system and lead to a reduction in the quality of official statistics.

5. CONDITIONS OF ACCESS

Only researchers belonging to a recognised research entity may request access to confidential data for scientific purposes. The research entity's duly designated representative is obligated to sign a confidentiality undertaking.

Access to confidential data for scientific purposes may be granted if:

- the research proposal submitted by the researcher(s) has been approved. Each research proposal must be countersigned by the contact person identified in the confidentiality undertaking;
- all researchers requesting access to confidential data for scientific purposes have signed a confidentiality declaration.

Researchers must keep microdata files secure to ensure it is not accessible by anyone who is not authorised to access the data.

Results of the statistical analysis that may contain information on individual respondents should also be kept secure.

The confidential data for scientific purposes must be stored on a password-protected computer. Access to the data must be restricted to authorised researchers explicitly named in the research proposal.

The intermediate results of analysis containing confidential data must be stored in a protected environment.

After expiry or completion of the project indicated in the research proposal (or in the event of termination of access by Eurostat), the principal researcher must destroy the dataset and any data or variables derived from it and sign a declaration to the effect that it has been ensured that all data have been destroyed. This obligation applies to the original data sent by Eurostat and to all derived data, except for the aggregated and/or analysed data as presented in the research results/reports.

6. DATASETS AVAILABLE

Confidential data for scientific purposes are available in two forms:

- "scientific-use files" - partially confidentialised data[2] delivered to researchers;
- "secure-use files" - available in Eurostat's safe centre in Luxembourg (non-confidentialised data).

Scientific-use files are available for the following data collections:

- European Community Household Panel (ECHP)
- Labour Force Survey (LFS)
- European Union Statistics on Income and Living Conditions (EU-SILC)
- Adult Education Survey (AES)
- Community Innovation Survey (CIS)
- Structure of Earnings Survey (SES)
- Information and Communications Technologies (ICT) usage by Households and Individuals
- Continuing Vocational Training Survey (CVTS)
- European Health Interview Survey (EHIS)
- European Road Freight Transport Survey (ERFT)
- Household Budget Survey (HBS)
- Harmonised European Time Use Surveys (HETUS)
- Farm statistics (FSS/IFS)
- EU Gender Based Violence Survey (EU-GBV)

Secure-use files are available in the Eurostat safe centre in Luxembourg or remotely from access points accredited by Eurostat for:

[2] Data to which special statistical disclosure control methods have been applied in order to reduce the risk of identification of the statistical unit(s) to an appropriate level and in accordance with current best practice.

- Community Innovation Survey (CIS)
- Structure of Earnings Survey (SES)

7. SPECIFIC RULES FOR PUBLICATION (SCIENTIFIC USE FILES)

When publishing the results of the statistical analysis for scientific purposes, researchers must comply with the specific rules laid down below.

7.1. Specific rules for European Community Household Panel (ECHP)

In all reports, including both published and unpublished papers, two thresholds related to cell size will be distinguished for ECHP cross-sectional results:

- below 20 observations (unweighted sample), results must not be published;
- from 20 to 49 observations (unweighted sample), results may be published but are to be individually identified (e.g. shown in brackets).

For confidentiality reasons, reports that include sample sizes must only mention 'less than 20 observations' and '20 to 49 observations' (i.e. not the actual number) for these two thresholds respectively.

For unweighted sample sizes below 20 observations, the actual number of observations may not be derived from (or combined with) other information available in the researcher’s reports, e.g. column or row totals.

The same rules apply for longitudinal results, except that the thresholds change to 10 (instead of 20) and to 30 (instead of 49) observations linked across time.

7.2. Specific rules for European Union Statistics on Income and Living Conditions (EU-SILC)

In all reports, including both published and unpublished papers, two thresholds related to cell size will be distinguished for EU-SILC results:

- below 20 observations (unweighted sample) or if non-response for the item concerned exceeds 50%, results must not be published;
- from 20 to 49 observations (unweighted sample) or if non-response for the item exceeds 20% and is lower than or equal to 50%, results may be published but are to be individually flagged (e.g. shown in brackets).

For confidentiality reasons, reports that include sample sizes must only mention 'less than 20 observations' and '20 to 49 observations' (i.e. not the actual number) for these two thresholds respectively.

For unweighted sample sizes below 20 observations, the actual number of observations must not be derived from (or combined with) other information available in the reports, e.g. column or row totals.

7.3. Specific rules for Adult Education Survey (AES)

Same as for EU-SILC (see item 7.2 above).

7.4. Specific rules for European Health Interview Survey (EHIS)

Same as for EU-SILC (see item 7.2 above).

7.5. Information and Communications Technologies (ICT) usage by Households and Individuals

Same as for EU-SILC (see item 7.2 above).

7.6. Specific rules for Household Budget Survey (HBS)

Same as for EU-SILC (see item 7.2 above).

7.7. Specific rules for Harmonised European Time Use Surveys (HETUS)

Same as for EU-SILC (see item 7.2 above).

7.8. Specific rules for Labour Force Survey (LFS)

In all reports, including both published and unpublished papers, three thresholds related to cell size will be distinguished for LFS results:

- Confidentiality threshold - below three observations (unweighted sample), results must not be published;

- Reliability thresholds - regarding reliability restrictions, Eurostat defines two limits, called 'a' and 'b'. Reliability limits depend on the sample size and design in the individual Member States. The limits are expressed in thousands of the weighted population. The sub-scenarios refer to the various sets of variables and related (sub)samples using different weights: quarterly: S_Q, annual: S_Y, biennial: S _2Y, household using individual weights: S_HH, household, using average weights: S_HHAVG and modules: S_MOD.

  Reliability limits for any group of countries should be calculated as the maximum of the values of the countries belonging to the group. With this approach, all country reliability criteria are fulfilled.

  - Estimates corresponding to a (weighted) population below limit 'a' must not be published;

  - Estimates corresponding to a (weighted) population between limit 'a' and limit 'b' may be published with a warning concerning their limited reliability. The limits vary across Member States, years, type of dataset (quarterly/yearly), and subsample. The thresholds 'a' and 'b' are provided in an Excel dataset available on the Eurostat website ([3]). More

([3]) https://ec.europa.eu/eurostat/documents/1978984/6037342/reliability_limits.xlsx

information can be found also at the LFS Statistics Explained pages, section "Publication guidelines and thresholds".

- Data on age in single years must not be disclosed in published tables. It may be published in broader age groups, such as the 5-year age bands provided in the derived variable AGE_GRP ([4]).

7.9. Specific rules for Community Innovation Survey (CIS)

Any statistics (tables, graphs, textual references) on any kind of sub-population cell must not be published if:

- they consist of less than 10 enterprises;
- one enterprise represents more than 70% of the total sub-population expenditures, employment or turnover;
- two enterprises represent more than 85% of the total sub-population expenditures, employment or turnover.

In addition, where there are primary confidential cells, secondary confidentiality methods must be applied to ensure that these primary confidentiality cells cannot be estimated with the help of the other non-confidential cells. For more information on how to apply confidentiality methods (statistical disclosure control), see chapter 6 of the manual: "How to use microdata properly. Self-study material for the users of European microdata".

7.10. Specific rules for Structure of Earnings Survey (SES)

Any produced output (tables, graphs, etc.) must not be extracted or published if:

- it consists of less than 10 enterprises / local units or employees;
- one enterprise represents more than 70% of the total sub-population employment or total earnings;
- two enterprises represent more than 85% of the total sub-population employment or total earnings.

In addition, where there are primary confidential cells, secondary confidentiality methods must be applied to ensure that these primary confidentiality cells cannot be estimated with the help of the other non-confidential cells. For more information on how to apply confidentiality methods (statistical disclosure control), see chapter 6 of the manual: "How to use microdata properly. Self-study material for the users of European microdata".

7.11. Specific rules for Continuing Vocational Training Survey (CVTS)

Any statistics (tables, graphs, textual references) on any kind of sub-population (cell) must not be published if:

([4]) LFS weighting factors are generally calibrated to 5-year age groups. This ensures reliability of results at this level.

- they consist of less than 10 enterprises;
- one enterprise represents more than 70% of the total sub-population employment or total labour costs;
- two enterprises represent more than 85% of the total sub-population employment or total labour costs.

In addition, where there are primary confidential cells, secondary confidentiality methods must be applied to ensure that these primary confidentiality cells cannot be estimated with the help of the other non-confidential cells. For more information on how to apply confidentiality methods (statistical disclosure control), see chapter 6 of the manual: "How to use microdata properly. Self-study material for the users of European microdata".

7.12. Specific rules for European Road Freight Transport Survey (ERFT)

The anonymised datasets may contain confidential information. The user of the data must therefore comply with restrictions related to the dissemination of tables described in Commission Regulation (EC) No 6/2003 of December 2002[5], Art. 3 (point 1):

Dissemination of tables to users other than the national authorities of Member States shall be subject to the condition that each cell shall be based on at least 10 vehicle records depending on the variable tabulated. Where a cell is based on fewer than 10 vehicle records, it shall be aggregated with other cells, or replaced with a suitable flag.

7.13. Specific rules for Farm Statistics (FSS/IFS)

In any tables or data to be disseminated the user will indicate a cell as confidential if:

- the (extrapolated) number of holdings that contribute to the cell (extrapolated) value is less than or equal to 4 and/or
- one or two (extrapolated) holdings represent more than 85% of the cell (extrapolated) value.

As a general rule, in any tables or data to be disseminated, the user will suppress a cell due to unreliability if the coefficient of variation of the cell estimate is equal to or greater than 35%. However, when computing counts or proportions, the unreliability criterion to use should be based on the standard error of the proportions (counts are first converted to proportions for this purpose[6]). If the standard error is equal to or greater than 0.175, the cell estimate should be suppressed.

The convention used is to replace a confidential value with ":c". Unreliable estimates are replaced with ":u".

For non-confidential and non-unreliable cells, the extrapolated number of holdings and all values of variables in cells are rounded to the closest multiple of 10.

[5] Commission Regulation (EC) No 6/2003 of 30 December 2002 concerning the dissemination of statistics on the carriage of goods by road, OJ L 1, 04/01/2003 p. 45 (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:001:0045:0049:EN:PDF)

[6] A count is converted to a proportion by considering the total number of holdings as denominator.

Because of the confidentiality treatment, the sum of the individual cells may not systematically match with the value of the "total" cell.

7.14. Specific rules for EU Gender Based Violence Survey (EU-GBV)

Following publication rules to be taken into account for dissemination results:

- estimates should not be published if based on fewer than 20 sample observations or if the non-response for the item concerned exceeds 50%; instead value ":u" (not available due to unreliability) should be used;
- estimates should be published with a flag "u" (unreliable) if based on 20 to 49 sample observations or if non-response for the item concerned exceeds 20% and is lower or equal to 50%;
- estimates shall be published in the normal way when based on 50 or more sample observations and non-response for the item concerned is lower than 20%.

8. INDICATION OF DATA SOURCES - EUROPEAN MICRODATA

The following statement should be used when providing information about source of the data used for research:

This study/report/paper is based on data from Eurostat, <name of the survey, reference year(s), release date, version and DOI reference if available>. The responsibility for all conclusions drawn from the data lies entirely with the author(s).

This formulation is specified in the Terms of use of confidential data for scientific purposes (heading: Identification of data sources).

The acknowledgment of data sources is mandatory and researchers commit to it by signing individual confidentiality declaration.

9. DATA MATCHING

Researchers are prohibited from attempting to link the microdata to other (including public) datasets, unless explicitly agreed by Eurostat. Matching two datasets increases the likelihood of the identification of statistical respondents represented in both datasets.

10. WHAT TO DO IN CASE OF A CHANGE IN THE SET-UP OF A RESEARCH PROJECT

Once a research proposal has been accepted, new researchers can be added (provided they sign a confidentiality declaration), and the duration of the project can be extended.

For more details how to make changes in the running projects, please refer to S**ection 7** of the guide How to apply for microdata access?

11. YOUR RESPONSIBILITY TO PROTECT CONFIDENTIAL DATA

Both the confidentiality undertaking signed by the duly designated representative of the research entity and the individual confidentiality declaration signed by the individual researcher provide the basis for legal action in cases where conditions of these documents have been neglected.

The Commission can take action in the event of a breach of confidentiality as follows:

- by revoking the offending researcher's (and if necessary, his/her research entity's) access to microdata;
- by suggesting the research entity takes disciplinary action against the researcher;
- by claiming civil law compensatory damages from the research entity. The confidentiality undertaking includes a reference to the applicable law and competent court;
- and/or by filing a complaint or by reporting the breach to the police on the basis of national legislation. The Commission may participate in national proceedings as plaintiff.

Depending on the situation, sanctions may be applied on researchers or their research entities.

12. LEGAL BASIS

Legal basis for granting access to confidential data can be found in the Regulation (EC) No 223/2009 on European Statistics[7] (Article 23) and in the Regulation (EU) No 557/2013 on access to confidential data for scientific purposes[8].

13. ANY QUESTIONS?

If any issue in this manual is unclear, or in case of further questions on the use of confidential data, please contact us: estat-microdata-access@ec.europa.eu.

[7] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:EN:PDF
[8] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF