EUROPEAN COMMISSION
EUROSTAT

Directorate G – Business and trade statistics
G.2 – European businesses

Community Innovation Survey (CIS) Access to microdata in the Safe Centre

Note to researchers visiting Eurostat's Safe Centre

Commission européenne, 2920 Luxembourg, LUXEMBOURG - Tel. +352 43011
Office: BECH B3/386 - Tel. direct line +352 4301-37468

Table of Contents

1. INTRODUCTION

Researchers receive the authorization to use microdata in the Safe Centre for an agreed research project and after fulfilling the requirements and procedures set up for this purpose: recognition of the research entity, approval of the research project and authorization of the concerned data providers.

Before entering the Safe Centre, the researcher must:

- get familiar with the data methodology and structure (all information can be found on Eurostat website)
- book the Safe Centre (several weeks before the visit);
- receive a user account for consultation of data.

2. EUROSTAT SAFE CENTRE

2.1 Visiting hours and assistance

The Safe Center can be accessed from 7 am to 8 pm. Assistance on IT infrastructure, statistical software or CIS data is assured during the office core hours from 9 am to 5 pm.

The first day of the visit a Eurostat responsible agent guides the researcher to the Safe Centre (an office in Eurostat building) and explains the procedures and modalities of access to the data.

2.2 Access to the data

Researchers work on a stand-alone PC that is connected to the repository dedicated to a given research project (RPP) via researchers' personal account. Researchers get access to the required data with a possibility to save all processing programs (ado files, scripts), methodology files, and all the outputs in a secure personal folder.

If another researcher wishes to work simultaneously on the same project (RPP) or if another researcher shall continue the previous researcher's work (e.g. to use the previous researcher's work files, data, programs and interim results…), this should be notified to Eurostat before the visit in order to arrange the working environment accordingly.

2.3 Statistical software

The statistical tools available in the Safe Centre are STATA15, R and Rstudio. Microsoft Office tools like Excel or Word are also available. Eurostat does not assist in data processing and does not provide support in using the statistical software.

The available statistical software contains the regular packages. If a researcher wishes to use a specific package (library), it must be sent to Eurostat before the visit. Only the non-standard specific packages of STATA (ado files etc.) and R can be uploaded by Eurostat on request. Researchers are in charge of the functionality of the packages.

2.4 Rules of use of Safe Centre rooms

In Safe Centre rooms it is NOT possible:

- to print documents;
- to copy data on external devices;
- to upload data/files from external devices;
- to copy data to the local hard disk;
- to connect recording devices to the serial, parallel and USB ports;
- to connect a laptop to the network;
- to connect to internet.

2.5 Researcher's own data

Linking own data set(s) with the datasets available in the Safe Centre is not permitted unless explicitly allowed in line with the approved research project proposal (RPP). The approved datasets must be sent to Eurostat before the visit.

2.6 CIS datasets

CIS datasets are available for different years and countries, see more here: https://ec.europa.eu/eurostat/documents/203647/771732/Datasets-availability-table.pdf

Researchers are granted access to the countries and years indicated in the approved research proposal (subject to availability, see link above). The data sets are provided in CSV format. Particular notes on the microdata are attached to the data.

3. OUTPUT VALIDATION

3.1 General rules

All results to be taken away from the Safe Centre must be checked for possible disclosure of confidential data. Disclosure checks cover both primary and secondary confidentiality.

The two main reasons for declaring data to be primary confidential are:

• Too few enterprises in a cell;

• Dominance of one or two enterprises in a cell.

Secondary confidentiality concerns data which are not primary disclosive, but whose dissemination, when combined with other data, leads to disclosure.

Safeguarding confidentiality is the responsibility of the researcher. The researcher should ensure that any results of the research intended to be published or otherwise disseminated do not contain information that may lead to the identification of persons or organisations represented in the data. Researchers shall be able to explain the processes to Eurostat agent and prove that the output is non-disclosive. Eurostat checks all output the researcher wishes to export from the Safe Centre.

Any deliberate attempt to compromise the confidentiality of persons or organisations to which confidential data for scientific purposes relates may result in prosecution in accordance with applicable law.

3.2 Specific rules for CIS

Any statistics (tables, graphs, textual references) on any kind of subpopulation **(cell)** will be rejected during output validation step:

(1) if they consist of **less than 10 enterprises**;

(2) where one enterprise represents **more than 70%** of the total sub-population expenditures, employment or turnover;

(3) where two enterprises represent **more than 85%** of the total sub-population expenditures, employment or turnover.

(4) if they contain maximum, minimum or quantile for the metric variables (turnover, number of employees, expenditures…)

Researchers should not propose for validation the statistics (tables, graphs, textual references) which do not comply with the rules 1-4 above. Eurostat will reject these statistics together with any other data whose dissemination could lead to disclosure (secondary confidentiality).

In the following sections there are few illustrative examples[1] of typical situations regarding confidentiality. These examples do not cover all possible situations. They show how the confidentiality rules shall be respected.

3.2.1 Example of primary confidentiality rule – minimum number of observations

Primary confidentiality means that any cell of the output to be exported from the Safe Centre needs to fulfil the conditions 1 to 4 above.

Table (1)

Region	Economic activity	Occupation	Median earnings	Number of local units	Number of employees
AA1	61	21	25.2	20	120
AA1	61	**22**	30.5	**4**	25
AA1	61	23	22.2	18	55
AA1	61	24	24.4	16	210
AA1	61	25	19.1	31	482
Total AA1	61	21-25	23.1	89	892

[1] The examples are drawn from another enterprise based dataset of Eurostat, namely the Structure of Earnings Survey data (SES).

Table (2)

	Variable B				
Variable A	Mod 1	Mod 2	Mod 3	Mod 4	Total
Cat 1	98	481	644	620	1843
Cat 2	12	**2**	38	**6**	**58**
Total	110	**483**	682	**626**	**1901**

In Table (1), occupation 22 does not fulfil the condition that the published cell shall have 10 or more units in it. This output proposal would be rejected as the number of local units in occupation 22 is only 4.

In Table (2), the cells [Cat 2;Mod 2] and [Cat 2;Mod 4] are below the threshold 10 which means that the output would be rejected.

3.2.2 Example of secondary confidentiality

Hiding occupation 22 in Table (1) or values 2 and 6 in Table (2) would create a problem of secondary confidentiality: a reader would be able to calculate the number of the local units in the hidden cell using the totals and non-hidden information. Also the hidden sensitive information, median earnings, could be easily estimated for occupation 22, at least its range. To avoid secondary confidentiality, tables 1 and 2 would be rejected entirely.

3.2.3 Example of dominance

In addition to the primary confidentiality rule of having at least 10 units in a published cell, the output must also fulfil the dominance rules 2 and 3.

Taking as an example the SES data, the dominance rules can be assessed for the number of employees in each unit and the gross earnings they represent.

Table (3)

Local unit/ Enterprise	**Average gross earnings in €**	Number of employees	Total gross earnings in €
1	2 333	5	11 665
2	3 535	603	2 131 605
3	2 802	10	28 020
4	2 956	12	35 472
5	1 999	6	11 994
6	2 716	10	27 160
7	2 350	9	21 150
8	2 752	20	55 040
9	2 232	6	13 392

10	1 998	10	19 980
Total 1-10	**3 409**	691	2 355 478

The whole Table (3) represents one cell with only total Average gross earnings € 3 409 aimed at publishing. However, as the second local unit / enterprise represents 87% of the employees and 90% of the gross earnings, the total Average gross earnings € 3 409 cannot be published. The dominance rule concerning the two largest units of a cell works similarly.

3.2.4 Regressions and other than tabular forms of output

Linear and non-linear estimation, simulations, modelling, different types of advanced analysis, particular indices and all (other) kind of econometric methods and their output may require a lot of specific knowledge to be able to validate the disclosiveness of the output.

In general, regression results are non-disclosive at an exact level (some inferences may be drawn within a margin of error in particular cases). Moreover, this small risk can be reduced further in ways which do not significantly reduce the usefulness of the results. The simplest way is non-reporting of incidental parameters, such as estimated constants or the coefficients on irrelevant dummy variables. In general, all regressions results are accepted if they are based on 10 enterprises or more.

For other analytical results, their disclosive nature depends on the manipulations carried out.

The assumption is that results are disclosive unless proved otherwise, and therefore it is in the researcher's interest to show that the results are non-disclosive. Graphs are treated as tables which just present the information in a different form.

Quantiles are also considered as tables. The same holds for maximum and minimum values with normally only one enterprise in a cell, i.e. they are confidential and cannot be released.

Detecting and protecting secondary confidentiality also for other than tabular forms of output shall be ensured.

3.3 Preparing the output for validation

Researchers must ensure compliance with the confidentiality rules before final submission of results for checking. Eurostat does not make proposals how to modify the output to get it accepted, but just indicates the conditions the output has to meet. If the output is rejected, it will not be sent to researchers. Researchers should hence check the results for disclosure of confidential data before leaving the Safe Center.

The proposed output shall include all information needed for output validation even if this (extra) information will not be published / used further (frequencies etc.). For example, the two last columns in Table (1) are necessary for validating the data even if they would not be published in the final report of the researcher. Note also that the two last columns would NOT be enough for Eurostat to validate the dominance rules (2-3) in Table (3). On the other hand, Table (3) includes the necessary information for validating the dominance rules. It is up to the researcher to decide on the type of presentation and the measures to show that all rules have been respected. Together with the results to be validated, all programs to derive this output will have to be presented as all results must be reproducible.

It should be noticed that the output validation concerns only the disclosiveness of the data. Output validation is not a quality check. Appropriateness of the assumptions or the underlying theory or analysis will not be assessed, nor the conclusions drawn. All this remains the researcher's responsibility.

It is in the researcher's own interest to provide the output which can be validated within reasonable time and without potential costly re-visits at the Safe Centre.

3.4 Output validation: form and time

The outputs intended to be taken from the Safe Centre need to be saved (together with the related programs) under dedicated directory. Table data shall be saved together with all other data necessary for the validation with the adequate headings, titles, and other metadata. New derived variables should be documented, and meaningful variable names be used. At the end of the research work, the researcher indicates to Eurostat the set of files to review. It is recommended that the researcher in person shows the output to Eurostat before leaving the Safe Centre and explains its main characteristics to facilitate the validation. Eurostat reserves itself the right to define the time needed for validating the output data. Eurostat will take all technical and organisational measures to ensure efficient checking without undue delays. Eurostat will make every effort to ensure that this delay does not exceed two weeks.

The validated output is sent to the researcher by email.

3.5 Rejection of output

The output is automatically rejected if the rules (1) to (4) above are not respected. The output may also be rejected if it is not fully understood or the output is very long. Same for non-documented or unexplained output. In these cases Eurostat cannot be sure whether the confidentiality rules are fully respected and cannot validate the data. The unaccepted output is simply rejected and not sent.

3.6 ACRO – automatic checking of research outputs

In order to optimize the process and reduce the burden of output checking, Eurostat developed the ACRO (Automated disclosure Control of Research Outputs) tool. These automatic procedures can be run with STATA selected statistical commands like ”tabulate” or “logit” and automatically mark the output as safe or not safe according to the rules presented in points 1 to 4 above. Statistics/commands not handled by ACRO must be separately submitted for review through the normal channel. Eurostat provides the researcher with the ACRO installation guide and explains the principles and advantages of the automatic checking system.

4. HOW TO CONTACT US

If you have questions or comments, contact us at ESTAT-STI-CIS@ec.europa.eu.