HUNGARIAN CENTRAL STATISTICAL OFFICE

# Guideline for researchers

December 2022

# CONTENT

# I. USE OF SAFE CENTER

## 1. INTRODUCTION – PURPOSE OF THE GUIDELINE FOR RESEARCHERS

The Safe Centre is a secure, camera-monitored environment, separate from the internal systems of the HCSO, which provides access to non-directly identifiable data sets for scientific purposes only, with a high level of protection of individual statistical data and strict compliance with data protection legislation.

The purpose of this document is to provide researchers with detailed information on the functioning of the Safe Centre of the HCSO and the rules they should follow when compiling research outputs in order to ensure effective data protection control and thus faster release of research results.

## 2. OPERATION/USE OF THE SAFE CENTRE

### 2.1. OPENING HOURS, APPOINTMENT BOOKING

**The HCSO Safe Centre can be used on working days from Monday to Thursday from 09:00 to 16:00 and on Fridays from 09:00 to 13:00.**

- **Entry to the Safe Centre is prohibited without an appointment!**
- **It is forbidden to be in the Safe Centre after opening hours !**

For the first scheduled research session, the Researcher should contact the Safe Centre Coordinator at kutatoszoba@ksh.hu.

At the agreed time, after the Researcher has received the access card at the main reception of HCSO (1024 Budapest, Keleti Károly u. 5-7.), the Safe Centre Coordinator will accompany the Researcher to the Safe Centre lobby and inform him/her **about the basic rules for entering the Safe Centre.**

After the first session, it is no longer necessary to arrange further research sessions with the Safe Centre Coordinator, but **an appointment must be made** before using the Safe Centre at the

**HCSO booking system.**

On the research day indicated in the booking system, the researcher will receive an access card at the main reception of the HCSO and will use the Safe Centre independently.

## 2.2. BEFORE ENTERING THE SAFE CENTRE

- **No personal items (bags, coats) are allowed in the Safe Centre!**

Before entering, personal belongings must therefore be placed in lockers designed for this purpose.

**Researchers are also not allowed:**

- **to bring in** or use communication and information recording devices (**telephone, laptop**) for any purpose,
- **to bring in or use any paper documents** (books, notes, etc.),
- **entering** the Safe Centre **without a valid research project**.

## 2.3. WORK IN THE SAFE CENTRE

On the first research day, the Coordinator will provide Researcher with the ID and password required to access the Safe Centre.

- **The password received must be changed by the Researcher at the first login!**

**Folder structure**

The Coordinator will show the Researcher the interface, the folder structure and check that the full set of requested data files has been outsourced.

- The data made available to the Researcher and the files the Researcher wishes to input will be placed in the ***"Bejövő"*** *inbox* **folder** of the Researcher interface. In the Inbox folder, the researcher has read-only access and cannot work on this interface.
- The analyses and their results can be placed in the ***"Kimenő"*** outbox **folder** according to the folder structure described in *Chapter II. 'Research results'*.

**Creating a note**

**Note taking is only possible on the whiteboards** (with markers also provided) in the Safe Centre **and in electronic format** (e.g. Word, Notepad) using the applications available in the research interface. The latter electronic document can also be requested by copying it into the appropriate folder in the Outgoing interface.

- **The note can only be taken out of the Safe Centre in electronic form, after a data protection check!**

Research results and **electronic notes** must be requested after leaving the Safe Centre by e-mail to kutatoszoba@ksh.hu.

During research work **in the Safe Centre it is prohibited**:

- **connecting** or attempting to connect **any device to client machines**,
- **copying** or attempting to copy **data files to any external media**,
- **intentionally concealing or damaging a camera,**
- damaging equipment in the Safe Centre,

**Research documentation**

As described in chapter II.2., it is necessary to prepare: the library structure expected in research work, the auxiliary tables for research results, the textual documentation of research results with or without program code. This document lists the file extensions that can be submitted for data protection verification. Furthermore, to request the outputs, it is necessary to fill in the so-called "Documentation form". The verification of the submitted outputs will not start until the form is incomplete or incorrectly filled in.

**Support for researchers**

On the first day of the research, the Safe Centre Coordinator will provide the researcher with information on how to apply for assistance.

If any questions or problems arise during the research, e.g. forgotten password, the researcher can contact the Safe Centre Coordinator on the following telephone numbers (also posted in the Safe Centre):

- 6621
- 6924

**Incidents, sanctions**

- If a Researcher who uses the Safe Centre violates the rules for the use of the Safe Centre, the HCSO is entitled to apply the following sanctions, depending on the nature of the violation:

- If the **breach is of an administrative nature** and did not seriously compromise the protection of the safe centre data - for example, the researcher did not register in the reservation system - **HCSO will issue a written warning to the Researcher**.
  After three consecutive warnings, which do not have to relate to the same research, HCSO has the right to suspend the researcher's access and ban the Researcher from using the Safe Centre at HCSO for 2 years.

- **If the breach of duty has led to a security incident** which, however, has not significantly compromised the protection of the data in the Safe Centre, **HCSO will record the breach of duty in writing**. **If a second security incident occurs**, HCSO is entitled to **ban the researcher from using the Safe Centre for 5 years**. Such a security incident shall in particular be deemed to be:
  - bringing in a communication, recording device
  - bringing in or taking out any written document (book, note, etc.)

- **If the breach of duty has caused a security incident which has significantly compromised the protection of the data in the Safe Centre**, **HCSO may, with immediate effect** and with the recording of the breach of duty, **permanently ban the researcher from using the Safe Centre** at HCSO. In particular, such a security incident is deemed to be:
    - the use of a communication, information recording and/or transmission device for any purpose
    - taking out one's own notes
    - removal of notes from the Safe Centre
    - use of the Internet, e-mail
    - connection of any device to a zero client
    - copying or attempting to copy data files to any external storage medium
    - deliberately covering up or damaging a camera
    - entering the Safe Centre without a valid research project

# II. RESEARCH RESULTS

- rules for the production of research results
- data protection rules, guideline

## 1. INTRODUCTION

The purpose of this chapter is to ensure that the role and operation of the data protection review of research results produced by the HCSO Safe Centre is transparent and as quick as possible for researchers using the Safe Centre service.

According to the Statistics Act (Act CLV of 2016 on Official Statistics)

> "The HCSO may (...) provide access for scientific purposes to individual data, the knowledge of which does not allow the direct identification of statistical units.
>
> (...) Only research results that have been verified by a member of the Official Statistical Service using appropriate methods prior to the release of the results, in order to minimise the risk of revealing statistical units in accordance with the best statistical methodology at the time, may be taken out of the secure environment."

HCSO may thus provide access to data deprived of direct identifiers (e.g. name, sort code) in the Safe Centre, but it must in any case check the research results produced in the Safe Centre environment from a data protection perspective before releasing them to the researcher and, if necessary, protect them by applying additional data protection measures.

The aim of output checking is to minimise the risk of revealing individual data[1] (e.g. data linked to a specific individual or company) from research results.

Effective data protection checking can be achieved if

- research results are transparent and properly documented;
- there is no need to apply data protection procedures, so that the research result can be released without modification because the risk of disclosure is minimal.

During the output checking activity, HCSO does not examine the professional findings of the research, the correctness of the mathematical-statistical solutions used, the professional content of the conclusions drawn; the examination is strictly limited to compliance with data protection aspects.

[1] Disclosure risk: the probability that at least one statistical unit can be identified from the published data or that previously unknown information about it will be disclosed.

Identification: the event where a statistical unit (in particular: natural person, enterprise, other institution) is clearly identified or one or more direct identifiers of the statistical unit are obtained.

Disclosure: the release of previously unknown information on a statistical unit from published data.

This guideline summarises the rules that the researcher must follow in order to have an effective data protection control. It is also in the researcher's interest to follow these rules, as it will allow a shorter time to check the results of the research and thus to have access to the results sooner.

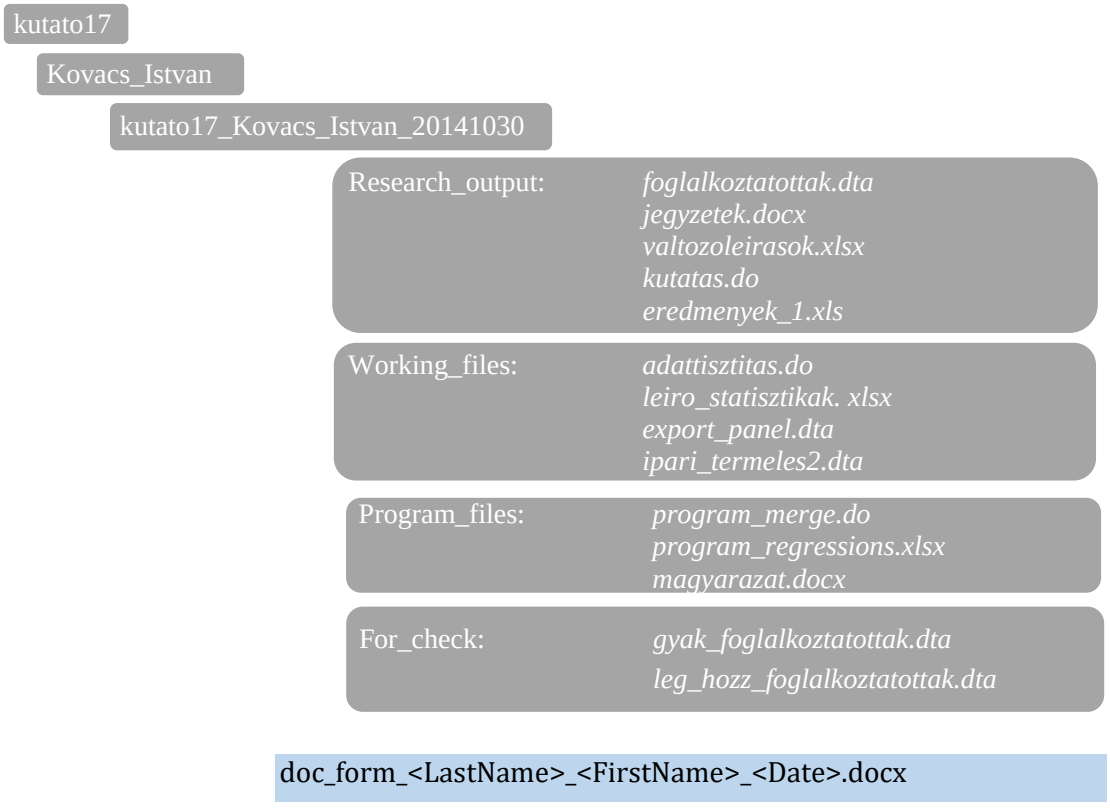
The guideline briefly discusses

- the requirements for research documentation,
- the auxiliary files and information to be produced for output checking
- the Statistical Disclosure Control methods

# 2. RESEARCH DOCUMENTATION

## 2.1 LIBRARY STRUCTURE EXPECTED IN RESEARCH WORK

The files required for data protection control should be stored in a dedicated repository for the research (*Kimenő\UserName*>) as follows.




<User name>: By user name we mean the research interface assigned to the research with a serial number (e.g. kutato17).

<LastName>_<FirstName>: The researcher should create a folder with his/her own name in the research interface (e.g. Kovacs_Istvan).

**<UserName>_<LastName>_<FirstName>_<Date>:** When the researcher wants to export a result, he/she creates a subfolder with the date of the day within the folder with his/her own

name (e.g. kutato17_Kovacs_Istvan_20141030). Within this folder, he/she creates the following four folders on the same level as each other:

- *Research_results*
- *Working_files*
- *Program_files*
- *For_check*

In addition, please save the documentation form with the following name **kut_eredm_dok_<username>_<lastname>_<Date>.docx**

1. **Research_output folder**: contains all the files that the researcher *wishes to export* (tables, regression results, program codes, text documents, etc.). These results are fully checked for data protection before they are released.

2. **Working_files folder**: contains all the files (working files) that were necessary to produce the files in 'Research_results', but which the researcher does not intend to export.

3. **Program_files folder**: contains all the program code and other related program code needed to produce the research results.

4. **For_check folder**: contains all files that assist in output checking. Please save the following in this folder:

- The number of observations associated with the research results, even if not required for the research. E.g. Typically, this is the case if proportions, distributions, percentages are included in the research result.
- If the results to be exported include magnitude table[2], then the auxiliary tables for the tables of the sum of the values:
  - *absolute frequency table;*
  - *table of the percentage of largest (dominant) contribution for each cell of the magnitude table.*
- In the case of graphs, the underlying data for the graphs.

For criteria for the construction of the auxiliary tables, see chapter 2.2.

Names of auxiliary tables: Please distinguish the file names from the original value sum table names by the prefix "**freq_**" or "**dom_**" (see *Annex 2*.)

[2] Tabular data: a set of micro-data that can be tabulated on the basis of the variables contained in the micro-data to produce a set of aggregated data. We distinguish between two types of tabular data, frequency tables and magnitude table

Magnitude table: a tabular data set whose cells contain the sum or average of the values of some criterion for individuals in a given sample or population with the characteristics that define that cell.

Frequency table: a tabular record whose cells contain the number of individuals of a given sample or complete set with a given set of characteristics.

Micro-data: micro-data are the primary form of data from which all other data are derived and in which they are stored.

## 2.2 PRODUCTION FOR AUXILIARY TABLES

If the results to be exported include a magnitude table, the researcher must produce auxiliary tables for that table for verification purposes: the *absolute frequency table* and the *share (percentage) of the largest contributor form the cell total* , also in a separate file. The **structure of the auxiliary tables should be the same as the original table**, with the same grouping criteria (Example. See *Annex 2*). The auxiliary tables **should be derived from the same micro-data** as the exported magnitude table.

- When calculating *absolute frequencies*, it is always necessary to check how many different respondents contribute non-missing values to the totals in the cells of magnitude table. (Respondents with zero values should be taken into account when determining the number of respondents. However, blank cells should be ignored.)

- *The table of the largest contributor's share in the magnitude table shall be produced according to the following formula:*

$$\frac{\max(|x_1|, |x_2|, .., |x_n|)}{\sum_{i=1}^{n} |x_i|} * 100$$

where x_i is the contribution of respondent i to the sum of the values of the given cell.

The auxiliary files ('feq_'; dom_') should be placed by the researcher in the *'For_check'* folder associated with the research results for that day.

## 2.3 DOCUMENTATION IN THE PROGRAM FILES

The documentation of the research (see *Annex 1*) within the programme files should be done taking into account the points listed below:

1. Please indicate at the beginning of the program file what the *purpose of the analysis* was, what files were worked on (directly or indirectly).
2. The logic and *structure of the analysis* (e.g. compilation of datasets, descriptive analysis, analytical results) *should be traceable* by including comments, explanations in the program file and by structuring the sections of the program file accordingly (title, short description).
3. The *renamed and newly created variables* are also defined in the program file in text form, with the researcher providing all the information relevant to the content of the variables.
4. If the researcher has used several program files in the research, it is advisable to prepare a "control" program file which calls the other program files in the appropriate order. Such a program file, annotated, can greatly aid transparency in more complex cases.
5. **For each output requested, please write an explanatory sentence on the Documentation form, this will make it easier to interpret the output, leading to faster checking and output release.**

Program files should be accompanied by documentation in English or Hungarian.

## 2.4 DOCUMENTATION OF RESEARCH RESULTS NOT PRODUCED BY PROGRAM CODE

Even if the research output is not produced by program code, the researcher must provide appropriate documentation for the submitted research output. The **process of generating the research result should be described in sufficient detail** to allow HCSO experts to recreate the research result (as he or she would be able to do from a program file). **All variables** used, renamed or **created must be properly and clearly defined**. The variables **should be labelled** according to the capabilities of the software used.

## 2.5 FORMAT OF RESEARCH RESULTS

Please convert your research results into one of the following extensions:

- tabular and other data[3] , results: xls, xlsx, xml, dta, acid, spv, sas7bdat, csv - *csv will be output in xls format if data protection modifications are required*
- Figures: ppt, jpg, bmp, png, pdf, gph - *editable gph figures can be requested only in justified cases with detailed description. In the folder "For_check", please also produce the underlying frequencies and values used for the figure.*
  *Figures are acceptable if the underlying data table does not contain low frequencies, individual data or if the value of the variable category cannot be read accurately from the axes.*
- note: doc, docx, txt, pdf - *include text description only, not tables highlighted from run results*
- syntax: sps, do, R, Rmd, sas
- log: smcl, log - *export only if no data protection is needed*

The same research result should not be exported in multiple formats, because if data protection is required on a given research result, exactly the same protection procedure should be applied to all file types. This is a time-consuming process for complex tabular data.

## 2.6 DOCUMENTATION FORM

If the documentation template is incomplete or has not been completed, the researcher will be asked to complete it. Output checking will not start until the documentation form is complete (see *Annex 4*)

[3] see footnote 2.

# 3. STATISTICAL DISCLOSURE CONTROL METHODS FOR OUTPUTS

The data protection rules for the output checking of research results are described below. The aim is to ensure that data protection considerations are already in place during the production of the research results and that researchers understand why the rules laid down for the Safe Centre are necessary. Before submitting your results for output checking, please review the contents of this chapter. If the rules described here are not met, it is advisable to re-examine the results and consider what needs to be modified or adapted to ensure that the output complies with all the rules mentioned below. Even if the submitted results have been produced in accordance with the criteria below, they still need to be checked by the experts of the HCSO, but the researcher will have access to the results to be exported after a shorter waiting period.

If the data protection rules are not taken into account by the researcher when producing the results, the HCSO will ensure data protection as described in chapter 3.1. If data protection is required, the tables produced may be restructured, too detailed categories may be merged or cell suppression (replacing them with the symbol "...") may be used.

## 3.1 DATA PROTECTION RULES FOR RESEARCH RESULTS

***General rules*:**

1. Prohibition to export microdata, microdata detail, tabular data with predominantly low absolute frequency cell values: no microdata or microdata detail may be exported from the Safe Centre. Also, tabular data with a vast majority of cells containing data from less than three reporting parties shall not be exported.
2. Threshold rule of three: According to rule of three, a cell is regarded confidential if the number of data providers is less than three[4].

***Additional rule for magnitude table:***

3. Dominance rule (1, k): in magnitude table, a cell is regarded confidential according to the dominance rule[5] (1, k) if the largest of the respondents contributing to the value exceeds k % of the value of the cell.

[4] The rule of three is not automatically applied to tabular data using population statistics. The possible relaxation of the rule of three is decided on the basis of the following criteria: how detailed is the spatial breakdown under consideration; what is the size of the population/sample; what is the sensitivity of the variables under consideration; how suitable are the variables under consideration for identifying observations (respondents). The former criteria are considered for variables such as: *geographical breakdown (e.g. county, municipality, enumeration area, etc.); classifications (e.g. HSCO, CPA, NACE, GFO, etc.); age; sex; marital status; country of birth; education; number of children in household; number of persons living in household; income; economic status; racial origin; nationality; health status; religion*. Researchers are advised to expect the rule of three to apply when submitting all research results.

[5] The value of the parameter k is not necessarily constant and may depend, for example, on the initial data and the level of detail of the results produced.

***Rules for statistical models***

*4.* Rules for model results: The export of model results (e.g. t-test, F-test, $\chi$2-test, regression results) requires that the model is based on a sufficient number of observations (min. 10) and that no model is run for a single firm/individual, even for time-series data.

If the research results to be exported do not comply with one of the rules described above, i.e. data protection methods need to be applied, the researcher or the HCSO expert may do the following:

— **restructure the table/merge categories that are too detailed**: reduce the number of sensitive cells, i.e. cells with values from 1-2 respondents, that are too numerous due to the use of too detailed categories by merging the categories. The restructuring must be done by the researcher, the expert can only suggest a restructuring of the table (see Chapter 3.3 and *Annex 3*.)

— **cell suppression**: If it is not possible to restructure the table, the cell values to be protected should be replaced by some convention (typically by "..."). This task is performed by the data protection expert (see Chapter 3.3 b); c) and *Annex 3*.)

## 3.2 DATA PROTECTION RULES CONSIDERED FOR EACH TYPE OF RESEARCH RESULT

| **Type of research output** | **Serial numbers of data protection rules according to the 4 basic cases given in chapter 3.1.** |
|---|---|
| Frequency tables | 1;2 |
| Magnitude tables | 1;2;3 |
| Maximum, minimum, quantiles | *2 |
| Mode | 2 |
| Average, index, ratios | 2;3 |
| Higher moments of the distribution (variance squared, covariance, shape indicators) | 2 |
| Graphs: visual representations of the data | ** |
| Linear regression coefficients | 4 |
| Non-linear regression coefficients | 4 |
| Summary and test statistics from estimates (t, F, R2, $\chi$2, etc.) | 4 |

*If the minimum and maximum values refer to a specific observation unit, they should be protected. To be released, it must be clear from the research documentation that the minimum and maximum values in the output do not refer to data from a specific observation unit.

** The exact value of any observation unit (e.g. maximum) should not appear as a data caption on or attached to the graph, e.g. as a background table to the graph.

### 3.3. PROPOSALS

A) *Restructuring of the table*. It is suggested that for variables such as *age, number of household members, number of children*, upper coding above certain values (e.g. age 85+, number of children 10+) or groups (e.g. age group 10) should be used).

B) *For multidimensional tables*, if a table is composed of three or more variables, at least one of which can be considered as a sensitive or ID variable, the likelihood of data protection intervention is increased, even at national level. Such variables could be e.g. *country of birth, year of birth, name of the field of highest completed education; nationality, NACE.*

C) *Multiple variable values*. In the case of variables with multiple categories, please select only those categories that are necessary for the research results (e.g. if only 10 of the 285 possible values of 'nationality' are sufficient for the research, only the narrowed down set should be used for the research results.)

D) *Differentiation*. Please note the data protection issues arising from the differences in the tables. If, for example, one table contains data on the total population and the other on the non-Roma population, the difference between the two tables will result in characteristics for the Roma population. In similar cases, please also produce the difference table.

## 4. EXTERNAL FILES

The researcher has the possibility to submit external files (notes, databases) to the Safe Centre, which can be used in the safe centre environment during the research. The inclusion of external files in the safe centre environment is not automatic; these files are also subject to a data protection check before they are made available. In this case, the checking is limited to examining whether or not the file contains direct identifiers. If so, the direct identifiers are removed (e.g. name, address, telephone number, sort code, etc.).

ATTENTION!

Must be provided for each file to be imported:

- the meta information of the variables in the file (variable names, exact content)
- a short description in Word format to help understand the content of the file.

## **ANNEX 1:** EXAMPLES FOR DOCUMENTING PROGRAM FILES

Example 1: Presentation of the research:

"In this research I analysed the productivity performance of manufacturing enterprises with more than 5 employees registered in Hungary, based on data from 2002. For this I used industrial production (iparitermeles.dta) and foreign trade (kulker.dta) data. At the beginning of the program file, data cleaning and the preparation and linking of the files were carried out. Afterwards, I filtered the manufacturing enterprises by NACE codes and ran linear regressions on them. In these I tested whether explanatory variables X, Y, Z significantly explained variable W. I also ran the model on different subsamples (*a, b, c*) and finally obtained the usual descriptive statistics for variables *u*, *v*, *h* and *e*.

Example 2: Graphical representation of the results

```
1  ************************************************************
2  ** (3) Az eredmények grafikus ábrázolása **
3  ************************************************************
4
5  /* Az egy főre jutó GDP növekedési üteme és az exportált autóalkatrészek
6     növekedési üteme közötti kapcsolat grafikus ábrázolása */
7  twoway (scatter GDP_ann_growth valcomp_ann_growth, msymbol(square) msize(tiny)
8          mcolor(black) mlabel(partnercode) mlabsize(tiny) mlabcolor(black))
```

Example 3: Calculating frequencies and maximum contributions to a table of value amounts

```
 1  ** Példa gyakoriságok és legnagyobb hozzájárulások kiszámítására**
 2  ********************************************************************
 3
 4  cd "D:\Kutatói_tájékoztató"
 5
 6  * 2014-es ipari termelés és értékesítés adatbázis
 7  use 2014, clear
 8
 9  * Összes értékesítés nettó árbevétele és átnevezések
10  egen ERT_ARBEV=rowtotal(IABB004 IABB005)
11  rename M005 megye
12  rename M0581 szakagazat
13
14  * Változók felcimkézése
15  label variable ERT_ARBEV "Értékesítés nettó árbevétele"
16  label variable megye "Megye"
17  label variable szakagazat "Szakágazat"
18
19  preserve
20  * Kutatási eredmény: Összes értékesítés nettó árbevétele
21  collapse (sum) ERT_ARBEV, by(megye szakagazat)
22
23  save kut_eredm_1, replace
24
25  * Adv_ell: Gyakorisági tábla
26  restore
27  preserve
28
29  collapse (count) ERT_ARBEV_gyak=ERT_ARBEV, by(megye szakagazat)
30
31  save gyak_kut_eredm_1, replace
```

```
32
33    * Adv_ell: Legnagyobb hozzájárulások
34    restore
35
36    gen ERT_ARBEV_abs=abs(ERT_ARBEV)
37    collapse (sum) ERT_ARBEV_sum=ERT_ARBEV_abs (max) ERT_ARBEV_max=ERT_ARBEV_abs,
38                   by(megye szakagazat)
39
40    gen ERT_ARBEV_lh=(ERT_ARBEV_max/ERT_ARBEV_sum)*100
41
42    drop *_max *_sum
43
44    save leg_hozz_kut_eredm_1, replace
```

## ANNEX 2.: EXAMPLE OF DATA PROTECTION CHECKS ON MAGNITUDE TABLES

If the research result is a value sum table, the following steps should be carried out:

**Table 1.: Result found in the 'Research_results' folder**

| Total sales turnover of enterprises by region and by section of the economy in 2010 (fictitious) | | | | | |
|---|---|---|---|---|---|
| Region | Sections of the economy examined | | | | |
| | A | B | C | D | Sum |
| 1 | 7 767 971 328 | 211 091 899 472 | 9 943 678 279 | 303 314 418 304 | 532 117 967 383 |
| 2 | 293 999 322 944 | 482 392 636 132 | 195 788 826 974 | 132 201 948 800 | 1 104 382 734 850 |
| 3 | 109 496 225 408 | 8 703 476 032 | 125 583 012 384 | 251 129 981 347 | 494 912 695 171 |
| 4 | 199 566 570 752 | 327 763 841 000 | 97 802 072 160 | 12 144 741 376 | 637 277 225 288 |
| 5 | 275 043 249 600 | 70 936 308 800 | 161 725 637 616 | 430 395 294 040 | 938 100 490 056 |
| 6 | 165 900 728 448 | 126 406 154 216 | 334 304 238 378 | 212 415 489 584 | 839 026 610 626 |
| 7 | 233 459 129 720 | 156 755 016 340 | 8 416 344 064 | 70 604 533 024 | 469 235 023 148 |
| Sum | 1 285 233 198 200 | 1 384 049 331 992 | 933 563 809 855 | 1 412 206 406 475 | 5 015 052 746 522 |

Step 1.: The researcher prepares and saves the absolute frequency table for Table 1 (Table 2) in the folder 'For_check'

When calculating the frequencies, the number of different respondents with non-missing values contributing to a given sum of values is taken into account. The auxiliary table should always be constructed according to the structure of the table of the sum of the values: e.g. if the table contains the turnover of enterprises broken down by NACE branch of activity, the frequency will be the number of different enterprises with non-missing values contributing to these turnover.

If in a value sum table there are value sums for several different variables (e.g. turnover, profit, value added, etc.), absolute frequencies must be compiled for each of them separately (e.g. because the number of missing values may be different for each variable).

**Table 2.: Absolute frequency table for Table 1**

| Number of companies by region and section of the economy in 2010 (fictitious) | | | | | |
|---|---|---|---|---|---|
| Region | Sections of the economy examined | | | | |
| | **A** | **B** | **C** | **D** | **Sum** |
| 1 | 1 | 42 | 10 | 54 | 107 |
| 2 | 51 | 97 | 40 | 26 | 214 |
| 3 | 19 | 2 | 30 | 57 | 108 |
| 4 | 35 | 69 | 19 | 2 | 125 |
| 5 | 51 | 15 | 34 | 79 | 179 |
| 6 | 33 | 28 | 70 | 43 | 174 |
| 7 | 43 | 35 | 1 | 14 | 93 |
| Sum | 233 | 288 | 204 | 275 | 1000 |

**Table 3.: Table of the largest contributors to Table 1 (%)**

| Share of the company contributing most to the value sum (%) | | | | | |
|---|---|---|---|---|---|
| Region | Sections of the economy examined | | | | |
| | A | B | C | D | Sum |
| 1 | 100.00 | 4.73 | 95.30 | 3.07 | 1.88 |
| 2 | 3.38 | 2.06 | 4.51 | 7.36 | 0.90 |
| 3 | 9.05 | 91.52 | 7.19 | 3.88 | 2.00 |
| 4 | 4.97 | 3.01 | 10.00 | 54.61 | 1.56 |
| 5 | 3.61 | 12.78 | 6.09 | 2.32 | 1.06 |
| 6 | 5.17 | 7.75 | 2.94 | 4.63 | 1.17 |
| 7 | 4.26 | 5.87 | 100.00 | 13.11 | 2.12 |
| Sum | 0.77 | 0.72 | 1.05 | 0.71 | 0.20 |

Step 2.: The researcher prepares and saves in the folder *For_check'* the table for Table 1 on the percentage share of the largest contributor in the value sums (Table 3)

Similarly, if there are value sums for several variables, the shares of the largest contributors for each of them should be prepared separately.

Step 3.: The researcher checks compliance with the data protection rules.

As can be seen in Tables 2 and 3, the rule of three is not always fulfilled, and for some cell values dominant contributors appear (using the (1.80)-dominance rule). Tables 4 and 5 highlight the cells to be protected (for Tables 2 and 3).

The protection of the selected cell values must be ensured, so the researcher can restructure the table or leave the research result unchanged, accepting that the values of the selected cells in the result table can be replaced by the symbol "...".

| Number of companies by region and section of the economy in 2010 (fictitious) | | | | | |
|---|---|---|---|---|---|
| Region | Sections of the economy examined | | | | |
| | **A** | **B** | **C** | **D** | **Sum** |
| 1 | 1 | 42 | 10 | 54 | 107 |
| 2 | 51 | 97 | 40 | 26 | 214 |
| 3 | 19 | 2 | 30 | 57 | 108 |
| 4 | 35 | 69 | 19 | 2 | 125 |
| 5 | 51 | 15 | 34 | 79 | 179 |
| 6 | 33 | 28 | 70 | 43 | 174 |
| 7 | 43 | 35 | 1 | 14 | 93 |
| Sum | 233 | 288 | 204 | 275 | 1000 |

| Share of the company contributing most to the value sum (%) | | | | | |
|---|---|---|---|---|---|
| Region | Sections of the economy examined | | | | |
| | A | B | C | D | Sum |
| 1 | 100.00 | 4.73 | 95.30 | 3.07 | 1.88 |
| 2 | 3.38 | 2.06 | 4.51 | 7.36 | 0.90 |
| 3 | 9.05 | 91.5 | 7.19 | 3.88 | 2.00 |
| 4 | 4.97 | 3.01 | 10.00 | 54.61 | 1.56 |
| 5 | 3.61 | 12.7 | 6.09 | 2.32 | 1.06 |
| 6 | 5.17 | 7.75 | 2.94 | 4.63 | 1.17 |
| 7 | 4.26 | 5.87 | 100.0 | 13.11 | 2.12 |
| Sum | 0.77 | 0.72 | 1.05 | 0.71 | 0.20 |

**Table 4. (top left): Data protection examination of an absolute frequency table**
**Table 5. (top right): Data protection examintation of a top contributors' table**

Step 4.: The data protection expert carries out the output checking

Assume that the expert:

- must perform a cell suppression check for Table 2 because of the rule of three (i.e. for the *light grey cells* in Table 4).
- must perform a cell suppression for Table 3 because of the dominance rule (1, 80) for contributors greater than 80% (i.e. for the *dark grey* cells in Table 5).

If we compare Tables 4 and 5, we can see that a total of 5 cell values need to be suppressed, due to the two rules mentioned above. This is shown by the '...' sign in Table 6.

**Table 6.: Cell suppression table**

| Total sales turnover of enterprises by region and by section of the economy in 2010 (fictitious) | | | | | |
|---|---|---|---|---|---|
| | Sections of the economy examined | | | | |
| Region | A | B | C | D | Sum |
| 1 | ... | 211 091 899 472 | ... | 303 314 418 304 | 532 117 967 383 |
| 2 | 293 999 322 944 | 482 392 636 132 | 195 788 826 974 | 132 201 948 800 | 1 104 382 734 850 |
| 3 | 109 496 225 408 | ... | 125 583 012 384 | 251 129 981 347 | 494 912 695 171 |
| 4 | 199 566 570 752 | 327 763 841 000 | 97 802 072 160 | ... | 637 277 225 288 |
| 5 | 275 043 249 600 | 70 936 308 800 | 161 725 637 616 | 430 395 294 040 | 938 100 490 056 |
| 6 | 165 900 728 448 | 126 406 154 216 | 334 304 238 378 | 212 415 489 584 | 839 026 610 626 |
| 7 | 233 459 129 720 | 156 755 016 340 | ... | 70 604 533 024 | 469 235 023 148 |
| Sum | 1 285 233 198 200 | 1 384 049 331 992 | 933 563 809 855 | 1 412 206 406 475 | 5 015 052 746 522 |

If we look further into the table, we can see that the table above is not yet protected, because for example the value of cell A1 can be calculated back from the Total. Thus, additional cells need to be suppressed, which is called secondary cell suppression. There are several solutions to this, one possible solution is shown in Table 7. This gives us the safe output that can be released.

**Table 7.: Publishable result**

| Total sales turnover of enterprises by region and by section of the economy in 2010 (fictitious) | | | | | |
|---|---|---|---|---|---|
| | Sections of the economy examined | | | | |
| Region | A | B | C | D | Sum |
| 1 | ... | 211 091 899 472 | ... | 303 314 418 304 | 532 117 967 383 |
| 2 | 293 999 322 944 | 482 392 636 132 | 195 788 826 974 | 132 201 948 800 | 1 104 382 734 850 |
| 3 | 109 496 225 408 | ... | 125 583 012 384 | ... | 494 912 695 171 |
| 4 | 199 566 570 752 | ... | 97 802 072 160 | ... | 637 277 225 288 |
| 5 | 275 043 249 600 | 70 936 308 800 | 161 725 637 616 | 430 395 294 040 | 938 100 490 056 |
| 6 | 165 900 728 448 | 126 406 154 216 | 334 304 238 378 | 212 415 489 584 | 839 026 610 626 |
| 7 | ... | 156 755 016 340 | ... | 70 604 533 024 | 469 235 023 148 |
| Sum | 1 285 233 198 200 | 1 384 049 331 992 | 933 563 809 855 | 1 412 206 406 475 | 5 015 052 746 522 |

If the row/column totals are not known and there is no realistic possibility that they can be obtained or calculated from other data sources, then the use of secondary cell suppression is not justified.

Step 5.: The researcher obtains the the safe output (Table 7).

## **ANNEX 3.:** EXAMPLE OF TABULAR DATA RESTRUCTURING

Suppose the researcher produces the following table. The table contains a detailed breakdown in terms of the number of children. When the researcher checks for compliance with the data protection rules, he sees that for some cell values the rule of three is not met. In this case, the researcher already knows that the low cell values should be protected, so he restructures the table.

**Table 8.: Resesarch results**

| | Number of children born alive in Zala county | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Family status** | **1 child** | **2 children** | **3 children** | **4 children** | **5 children** | **6 children** | **7 children** | **8 children** | **9 children** | **10 children** |
| unmarried | 17 | 3 | 3 | 3 | | 1 | | | | |
| married | 54 | 27 | 15 | 11 | 5 | 1 | 1 | 2 | | |
| widow | 71 | 28 | 14 | 9 | 1 | 2 | | | 1 | 1 |
| divorced | 11 | 4 | 4 | 5 | | | | | | |

The restructuring of Table 8 is shown in Table 9. By upper coding the number of children, we obtain a table that is adequate from a data protection point of view. The resulting Table 9 is forwarded by the researcher for output checking.

**Table 9.: Restructured tablet**

| | Number of children born alive, Zala county | | | |
|---|---|---|---|---|
| **Family status** | **1 child** | **2 children** | **3 children** | **4+ children** |
| unmarried | 17 | 3 | 3 | 4 |
| married | 54 | 27 | 15 | 20 |
| widow | 71 | 28 | 14 | 14 |
| divorced | 11 | 4 | 4 | 5 |

## **ANNEX 4.** DOCUMENTATION FORM

# Documentation Form of the Research Output

## 1. Basic research information

| | |
|---|---|
| **1.1** Researcher login ID: | |
| **1.2** Name of research project: | |
| **1.3** Name of research institution (name of organisation): | |
| **1.4** Name of the researcher who produced the research result: | |
| **1.5** E-mail address of the researcher producing the research result: | |
| **1.6** Date of submission of research results (dd/mm/yyyy): | |

A necessary condition for the export of research results from the secure environment of the HCSO is the correct completion of the research results documentation, which is checked by the coordinators of the department responsible for the user relations, and the researcher is contacted in case of problems or incorrect completion. **Only correctly and completely filled in forms will be forwarded to the experts carrying out the output checking of the research results.**

## 2. Documentation of research results

**Responses are mandatory, please read carefully the instructions for completing the spreadsheet.**

Please note that the format of the research results file can only be as below:

- tabular and other data and results: xls, xlsx, xml, dta, acid, spv, sas7bdat, csv - *csv will be provided in xls format if data protection modifications are required*
- Figures: ppt, jpg, bmp, png, pdf, gph - *editable gph figures can be requested only in justified cases, with detailed description. In the folder "For_check", please also include the underlying frequencies and the production of the values used for the figure.*
- *Graphs are acceptable if the underlying data table does not contain low frequencies, individual data or if the value of the variable category cannot be read off the axes accurately.*
- Note: doc, docx, txt, pdf - *include only your own thoughts, not tables highlighted from run results*
- syntax: sps, do, R, Rmd, sas
- log: smcl, log - *export only if no data protection is needed*

**Please make sure that you have prepared and copied into the folder "For_check" the auxiliary tables for your research results that are necessary for the output checking**: frequency tables for the value sum tables and the figures, and tables for the largest contributor's share (see chapter 2.2. Research results)

**Completion aid:**

**Name of output**: please list all results with extension that you copied to the "Research results" folder.

**Type/production method of output**: e.g. frequency table, magnitude table, regression analysis, log file. If the research result is a log file, it is not necessary to list all the tables, regression outputs, etc. in the log file, but please include a short comment or description of the results in the program file. **Log files should only be provided if no data protection is required!**

**Program file name and lines**: Enter the name of the program file and the lines that produce the research output. If you have produced the result **without program code**, please leave this column blank, but in section 1.2, briefly describe the purpose of producing the research result, the logical flow of producing the research result, the main steps of its production and provide the definition of the newly formed or renamed variables in Table 2.

**Input datasets used in the production of the result (with years):** for HCSO inputs, e.g. Census (2011), Mortality (2009), R&D data (2017-19), etc. For external datasets imported by the researcher, it is also necessary to indicate their exact name and, if not clear, the period.

**Explanation of content and context**: please provide a sentence explaining the content of the research result e.g. "Fetal loss rates at national level", "OLS regressions with fetal loss rates for high educated vs. low educated women", "the data file contains descriptive data on the size of industry clusters and the share of foreign firms".

Also indicate whether the research output is related to your current or previous output e.g.: Previously you have analysed/explored data at the municipality level, now at the county level. Previously, you only analysed companies with more than 250 employees, now you analyse the whole range of companies.

## 1.1 Research results produced with program code

Table 1.

| **Name of output** | **Type/production method of output** | **Program file name and lines** For production without program code, completion of chapter 1.2 is mandatory. | **Input dataset** used to produce the result | **Content explanation and context** |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

## 1.2 Research results produced without program code

Please give a brief description of the **purpose** of the research, the **logical process** and the **main steps** of the research!

**Please provide the definition of the newly created or renamed variables without program code!**

Table 2.

| Name of variable | Meaning of variable | Which file is the variable in? |
|---|---|---|
| | | |
| | | |
| | | |
| | | |