Guideline for researchers

Instructions about conducting research using the Safe Centre of the Hungarian Central Statistical Office

Hungarian Central
Statistical Office

July 2014

Table of Contents

Aim of the guideline

The purpose of this document is to give researchers information about legal, administrative, technical and methodological questions concerning access to the Safe Centre of the Hungarian Central Statistical Office (HCSO). The guidelines include the conditions of this safe environment, focusing on the requirements researchers' output has to meet, and on the rules of output checking.

The document aims to facilitate and speed up the process of output checking by helping both the researchers and Statistical Disclosure Control (SDC) experts. Moreover, the document provides useful information for those interested in the channels of data access available at the HCSO and it also informs them about the process and rules of output checking.

1. Process of the research and legal background of Statistical Disclosure Control

The Hungarian Central Statistical Office collects and manages data only for statistical purposes, it does not transmit individual data to third parties, except to provisions set by legal acts. In order to fulfill each legal requirement that concerns individual data and to retain the trust of the data providers, the HCSO gives special attention to the protection of individual data. In accordance with its obligation to provide data, the statistical data managed by the HCSO is used for dissemination and to support scientific research.

The HCSO provides several channels for the access of statistical data. Some of these channels are dedicated to provide access to researchers to prepared statistical data. These are called channels for scientific purposes. In this case, apart from the legal and methodological actions taken to ensure the protection of individual data, the physical data protection also has a prominent role.

Data access for scientific purposes can be granted via the Safe Centre, remote access, remote execution and release of anonymised microdata sets. These guidelines focus on the Safe Centre access. The Safe Centre is a strictly monitored environment within the premises of the HCSO, which is separated from the internal network of the HCSO and all the external network connections (internet). The legal obligations regarding the protection of individual data is fully applicable to the research outputs produced in the Safe Centre.

1.1. Process of the research

The following figure demonstrates the main steps of the research process in the Safe Centre, starting from the submission of the "Data request form for Safe Centre access" to the release of the research output after the output checking has been executed by the SDC experts.

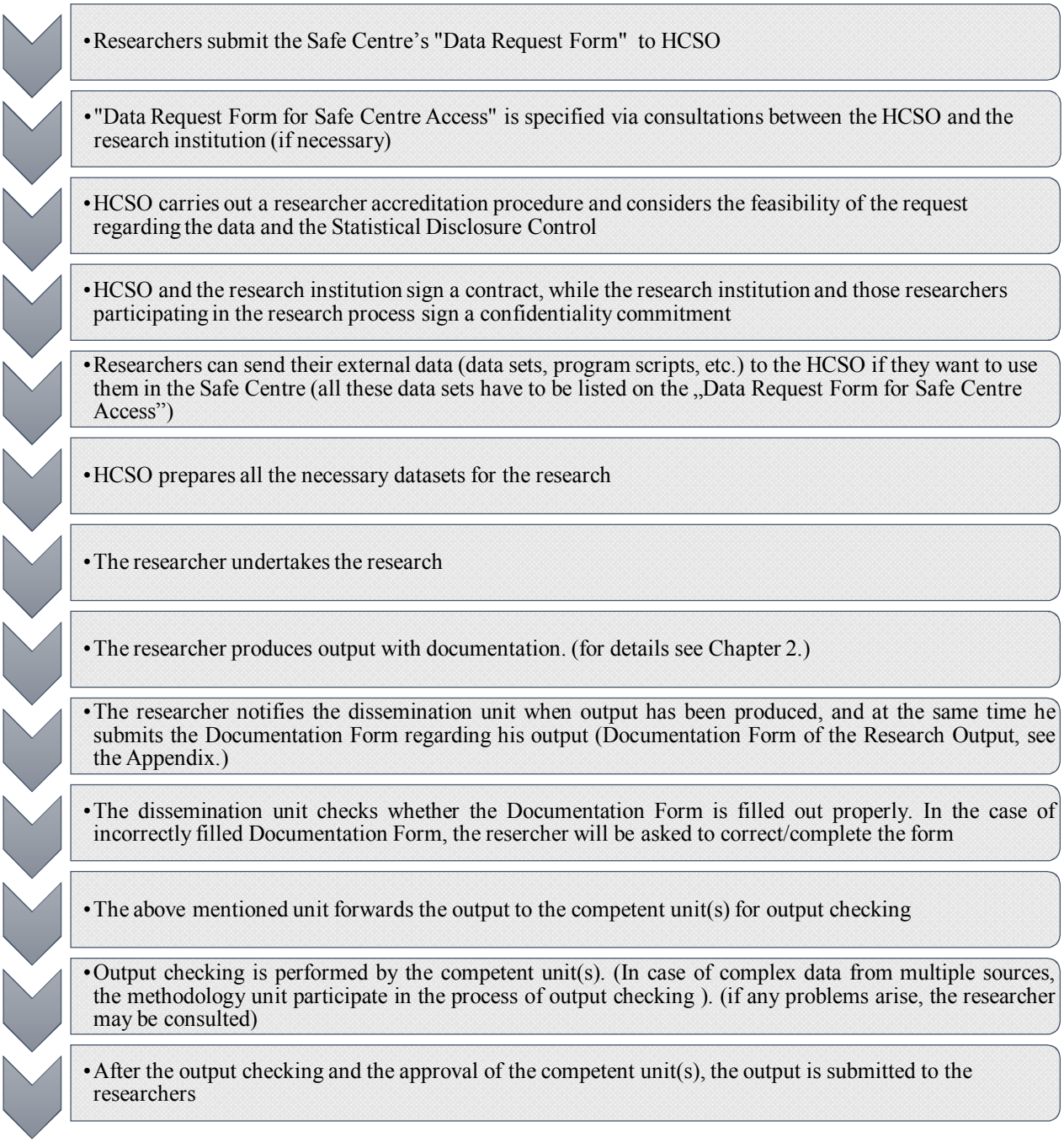

Figure 1. The process of the microdata access in a safe environment

We distinguish two kinds of data access regarding the output checking: simple and complex data requests. Based on this classification, the research can be carried out using:

- only one data owner unit's microdata[1]. In this case the data request is called **simple data request** while the output of the research is called output resulting from a simple data request. The output checking and the approval are made by the data owner unit whose data is used for the research.
- microdata[2] belonging to more than one data owner unit. Here we talk about a **complex data request** and output resulting from a complex data request.
  The methodology unit performs the output checking (it gives a professional statement), while access to the output has to be granted by all data owner units.

In order to have a smooth process (which was depicted on Figure 1.):

- ✓ the researcher should unambiguously declare his intention to take out his output to the dissemination unit, and he also needs to give information about the folders where his Documentation Form can be found.
- ✓ the researcher follows the instructions given by the HCSO regarding the folders' hierarchy when storing his working files and output (see section 2.1)
- ✓ according to this Researcher's Guidelines, the researcher creates all the „auxiliary tables"

If the output made by the researcher can only be released with high information loss, the HCSO informs him of that and recommends the preparation of the research output using alternative methods or aggregating the data on a higher level.

1.2 Tasks of the expert[3] responsible for output checking

The main task of the expert is to ensure that output is adequately protected against disclosure, that is to perform the output checking. We emphasize that the aim of the output checking is not to check the correctness of the methodology used in the research, but to decide whether the output fulfills all legal[4] and methodological requirements regarding SDC.

1

Microdata: record-level dataset containing information on observation units. The observation unit can be a person, a household, a dwelling, an economic entity, a transaction, etc.

[3] SDC expert or expert

[4] See Chapter 1.4.

The expert only inspects the output with respect to data protection. Neither the SDC expert nor the HCSO are responsible for or can be obliged to verify or refute the correctness of the output.

The duties of the *SDC expert* are the following:

- Checking all the external datasets sent by the against disclosure (input checking)
- Examining all the output produced by the researcher with respect to SDC, stating his/her opinion in a written form as to whether the output can be released in the current form
- Performing SDC methods on the research output to protect it if the risk of disclosure is high
- Making the output available to the dissemination unit
- Consulting the researcher if any questions arise regarding the SDC methods

1.3 Tasks of the researcher

The planning of the research and the choice of the appropriate methods used throughout the research are the researcher's responsibilities, thus the only person exclusively accountable for the professional content of the published output is the researcher[5] himself.

The researcher must respect the rules of the Safe Centre which can be found on the HCSO's website at http://www.ksh.hu/safe_centre_access. In addition, he/she has to facilitate the process of output checking by adhering to the following rules:

- For each external dataset that he wants to bring into the Safe Centre, he provides the variables' metadata (name of the variables, their exact content and as well as any descriptions which may help the HCSO to understand the external dataset)
- Acknowledging that the HCSO does not make available any datasets in the Safe Centre that contain direct identifier(s)
- For each magnitude table he wants to take away, he produces the respective frequency table in a separate file (See Table 3.)
- For each cell in a magnitude table, he calculates the share of the largest contribution from the total cell value (% share) in a separate file (See Table 4.)
- Researcher defines and labels all the created or renamed variables (See section 2.2)
- Researcher provides documentation and explanations for the output (See section 2.2)

[5] The HCSO does not intervene in the research concerning the profession, the researchers have a green light.

If the output checking requires, the SDC experts will ask the researcher to provide more explication regarding the output in either written or oral form, and to complete the documentation and the applied methods' descriptions.

1.4 Legal background

At HCSO, the collection of statistical data is based on the data providers' trust. This is why the HCSO considers the protection of individual data very important and ensures that it is protected in every stage of the research process at the Safe Centre. All the information regarding the legal background is available on the website of HCSO at http://www.ksh.hu/information_on_confidentialty_for_data_providers.

2. Documentation of the research

Understanding the research output is necessary to check it, therefore the researcher is obliged to do the documentation of the research according to this chapter. As long as the documentation is incomplete or is not carried out according to the description, the output checking cannot be executed. In this case researcher will be asked to complete the documentation.

2.1 Structure of directory for storing files

In location dedicated to the research (*X:\Kimeno\<Username>*) datasets necessary for the output checking must be stored as follows:

- HCSO assigns a research interface with a line number to each of the researches before the beginning of the research process (E.g. 'research17' will be assigned to the given research).
- Every researcher participating in this research can log in with the same username (this is the name of the research interface, e.g. 'research17') and password under the research. One research interface can be used by only one researcher at the same time.
- Thus on a research interface the researchers of the same research can see, modify and even delete the works of each other.
- Researcher is obliged to create a folder named after his/her own name (e.g. John_Smith) on the interface assigned to the research.
- Every time the researcher would like to take out the research output, he creates a directory named '<Username>_<Forename>_<Surname>_<Date>'. <Username> is replaced by name of research interface, <Forename>_<Surname> is replaced with the

forename and surname of the researcher (without accents) who handed in the output, <Date> is replaced by the date of handing in the output in order of day, month, year (without deliminators).

- In this directory the researcher creates three folders for the research output and the files related to the output named as follows: *'Research_Output'*; *'Working_files'*; *'For_check'*.

Example:

Assume that John Smith who participates in the research on the interface *research17* prepared an output that is submitted by him for output checking on 30th October 2014. The researcher must create a directory named *'research17_John_Smith_30102014'* in folder *research17*. In this directory he must create at the same level subfolders named *'Research_Output', 'Working_files'* and *'For_check'*

- The above mentioned directories' hierarchy looks like the following: *research17*
  - *John_Smith*
    - *research17_John_Smith_30102014*
      - *For_check*
      - *Research_Output*
      - *Working_files*
      - *output_doc_John_Smith_30102014.docx*

It should be emphasized that output checking will be carried out on the basis of files stored in the directory defined above. The researcher is free to create also other directories to store files, but since that files will not be checked, cannot be disseminated as well.

- Directory '**<Username> <Forename> <Surname> <Date>**' contains all of the files were used by the researcher for creating output submitted on the given date and all of the output files he would like to take out of the Safe Centre environment (syntaxes, working files, files necessary for output checking and also the output he would like to take out and the Documentation Form)
- Directory '**Research_Output**' contains all of the files the researcher *would like to take out* (tables, regression output, syntaxes, text documents, etc.)
- Directory '**Working_files**' is to store all of the working datasets the researcher *would not like to take out* but were necessary to prepare the files located in folder

'Research_Output' (From which working dataset an output origins must be indicated on the output Documentation Form.). Syntaxes that generate working datasets (the inputs of the process of generating results) from raw datasets provided by the HCSO that the researcher can access and/or from the external datasets submitted by the researcher are also located in folder *'Working_files'*.

– Directory *'For_check'* contains the follows:
  - *frequency table contains unweighted cell counts* for each magnitude table.
    The frequency table must be named as the magnitude table with prefix '**freq_**'.
  - table of the percentage of largest (dominant) contribution for each cell of the magnitude table.[6]
    This table must be named as the magnitude table with prefix '**dom_**'.
  - syntaxes used directly for preparing the output and the related syntaxes
  - definitions of variables occur in output stored in a transparent way (definitions can be given in comments of a syntax, text documentation, Excel spreadsheets, etc. in addition variables must be labelled in datasets. The aim is that all of information needed for output checking to be available and the definitions of variables and the ways of generating them to be found)
  - all the rest documents and tables that can promote the output checking (e.g. descriptions about the purpose, process and structure of the given research project, the order of generation of output, logs, etc.)

Output checking can only start after the researcher completely filled out the Documentation Form (see Annex). Documentation Form of the Research Output must be named as *output_doc<Forename>_<Surname>_<Date>.docx*

where *<Forename> and <Surname>* obviously must be replaced by forename and surname of the researcher and *<Date>* must be replaced by date of submitting output as described previously.

Example:

➔ *research17*
  - *John_Smith*
    - *research17_John_Smith_30102014*

[6] For every cell the share of the largest contributor in the value of the cell in percentages must be given.

- *Research_Output*
  - data_prepared.dta
  - balance_output.sav
  - notes.docx
  - description_of_variable.xlsx
  - research.do
  - results_1.smcl
  - …
- *Working_files*
  - clear_data.do
  - descriptive_statistics.sps
  - export_panel.dta
  - industrial_production2.sav
  - …
- *For_check*
  - freq_data_prepared.dta
  - freq_ balance_output.sav
  - …
  - dom_ data_prepared.dta
  - dom_ balance_output.sav
  - …
  - program_merge.do
  - program_regressions.sps
  - program_merge.smcl
  - ...
  - *explanation.docx*
- output_doc_John_Smith30102014.docx

2.2 Documentation in syntaxes, labelling variables

Correct documentation of the research process and output make the outputs easier to follow and to be understood by the SDC experts, which significantly decreases time of output checking. Documentation of research in syntaxes (scripts) required to do as follows:

1. Syntax must be commented in a proper quantity. *Logic and structure of the analysis must be easy to follow* (e.g. prepare datasets, descriptive analysis, analysis results).

   Please place particular emphasis on documenting the parts of syntax that are related directly to the output or indirectly but closely to the output.

   At the beginning of the syntax it is required to indicate which HCSO's datasets the researcher actually used (directly or indirectly).

   At beginning of syntax it is also required to give a brief description on *the aim of the analysis*, what kind of datasets were used to the work, what were the results (see Example 1) and what were the main steps during the preparation of the results.

   This information will show the main parts and structure of the syntax and give us some kind of context of the results that will thus be easier to understand.

   Example 1:

„In this research I analysed the productivity of the Hungarian manufacturing companies have more than 5 employees in 2002. I used *balance sheet data* (balance_sheet.dta) and *trade data* (trade.dta). At the beginning of the syntax, I performed *data cleaning* then I merged the two datasets. I filtered out companies operating in manufacturing industry by NACE codes, then run a *linear regression* model on them. In this regression I examined if the respondent variable *W* estimated significantly by the explanatory variables *X, Y* and *Z*. The model was run also on subsamples (*a, b, c*), and finally for variables *u, v, h* and *e* descriptive statistics were also made."

2. Main parts of the syntax must be clearly separated. For each part of the syntax, there must be a brief description, a title must be given to each part, to clarify what is happening in the concerned part and to clearly indicate which output files are prepared from which part of the syntax.

   Example 2:

```
**************************

** (1) Checking consistency of the database **

**************************
```

/ Checking consistency of the variables 'id', 'birthdate' and 'gender' */*

cd X:\Kimeno\researcher17\John_Smith
use labourforce_database1, clear

replace birthdate=birthdate[_n-1] if (id==id[_n-1])
replace gender=gender[_n-1] if (id==id[_n-1])
…

****

** (5) Graphs **

****

* 'Old' member countries *

****

/ Creating scatter plot to demonstrate the relationship between growth rate of GDP per capita and growth of the exported vehicle components */*

twoway (scatter GDP_ann_growth valcomp_ann_growth, mysymbol(square) msize(tiny) mcolor(black) mlabel(partnercode) mlabsize(tiny) mlabcolor(black)) ///

(lfit GDP_ann_growth valcomp_ann_growth) if country_dummy==0, ytitle(Annual average for growth rate of GDP per capita (%))ytitle(, size(small)) ///

ylabel(, labsize(small)) xtitle (Average annual growth rate of the exported vehicle components (%)) xtitle(, size(small)) title(Relationship between the growth rate of GDP per capita and the exported vehicle components (Old member countries), size(small)) ///

note((GDP per capita at 2005 constant prices; Vehicle components export at 2006 constant prices), size (small) position(6)) legend(size(small))

```
...
```

3. Definition of renamed and newly created variables in the syntax, by which all information is given concerning the content of these variables.

   Example 3:

```
**********************************************************
** (1) Rename variables, generate new variables **
**********************************************************

/* Rename balance sheet variables */

ren Shorttermliabilities s_term_debt
ren Accountspayable payables

/* generate variable for short-term credit */

gen scredit=s_term_debt-payables

...

**********************************************************
** (2) Label variables **
**********************************************************

/* Creating two contracted categories for vehicle export: component (0) and final vehicle
(1) */

label define vehicletype 0 "components" 1 "final vehicles"

label values comm_code vehicletype

label variable comm_code "Category var.: Components and final vehicles"
```

4. If numerous syntaxes were used in the course of the research and for the preparation of the output, it can be useful making a 'master code' calls other codes in a proper sequence. In complicated cases a commented master code can be helpful in understanding the research process.

   Example 4:

```
**************************************************************************

/* This syntax creates output and the respective auxiliary tables from the original datasets by
calling other syntaxes.*/

**************************************************************************

cd "X:\Kimeno\researcher17\John_Smith\"

/* Preparing panel dataset by cleaning data and merging trade and balance sheet datasets */

do prepare_paneldataset.do

...

cd "X:\OUT\researcher17\John_Smith\Check"

/* Aggregation for years, this syntax creates the output and the auxiliary tables */

do aggreg_year.do
```

2.3 Other possibilities for research documentation

Researcher is obliged to provide adequate documentation for the output even if it was not generated by a program file. In this case researcher is free to choose a convenient way for documentation in order to make the aim and the process of the research transparent and to provide all the rest information needed for the output checking. Description of the process of creating output must be so detailed that on the basis of this, the SDC expert checking the output is able to recreate the output (as the output checker could recreate it using a syntax). All variables used, renamed or generated by the researcher must be precisely defined. Variables must be labelled as far as possible in the software used.

Of course, files that help understanding the output can be useful in every case, but if the output is not generated by syntax then providing these kind of files is not only useful but also obligatory.

Examples for files that help output checking:

text documents: descriptions of the aim and structure of the research and about the process resulting in producing outputs.

tables: it can be a suitable form e.g. to collect definitions of variables and to follow up renaming variables.

make software to display syntax: e.g. in case of using SPSS menu there is a possibility to click 'Paste' button instead of 'OK' that would mean execute. 'Paste' pastes syntax of commands generated by the software in a program file.

turn on logging, save log files: log files can help the SDC expert to understand the process of the research.

3. Methodological procedures and rules of output checking

3.1. The applied sensitivity rules in the HCSO

In the course of output checking several sensitivity rules[7] can be applied[8]. Hereafter only rules adopted by the HCSO in its current practices are outlined.

minimum frequency rule: A cell of a table is considered to be unsafe if the number of the respondents is below a given threshold.

> Remark: For instance, let the threshold *m* equal 5. The cell value is considered to be confidential if the number of respondents is less than 5.
>
> ***threshold rule of three***: This is a specific threshold determined by the enforcement decree of the Hungarian Statistical Act. According to the decree it is not permitted to disseminate any information – regardless if data is on individual level or aggregated level – where the number of respondents is less than 3 (Decree of the Hungarian Statistical Act, 19. §).
>
> Remark: Application of stricter minimum frequency rule (m>3) is allowed, but the breach of the threshold rule of three (m<3) is prohibited. Taking into account at least the threshold rule of three ($m \geq 3$) is not only justified by the legal regulation, but this is the lowest threshold which prevents the *unambiguous* recalculation of the number of respondents for a protected/suppressed cell. The probability of the *identification* and/or *disclosure* of any respondent's attributes is reduced in that way. The *tabular output* must meet at least the requirement of the threshold rule of three. The minimum frequency rule is considered also for such magnitude tables where the respective frequency count tables are not requested to be taken out as researcher output.

[7] E.g.: minimum frequency rule, dominance rule, p% rule, p/q rule

[8] *Handbook on Statistical Disclosure Control* (http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf).

(n,k)-dominance rule: A cell is regarded confidential, if the *n* largest units contribute more than *k*% to the cell total.

> Remark: The value of "*n*" and "*k*" are set taking the professional opinion of the data owner HCSO unit into consideration. Their values can vary from output to output depending on the actual disclosure risk. The researchers are informed about the applied SDC methods and sensitivity rules, i.e. minimum frequency rule and dominance rule which make the research output safe. However, exact values of the parameters must not be released.
>
> Dominance rule is taken into account in case of *magnitude tables*. The application of the dominance rule could be justified if a research project pertains to high *concentrated* industries where low number of so called dominant firms cover the vast majority of the market share. In that case disclosure of any industry characteristic would result in quite a precise estimation on a given attribution of the *dominant firms*. Moreover, if a low number of firms operates in a given industry, there is a risk that any attribution of the dominant firm is set by means of constituting *coalition* of the other *(n-1)* firms.

3.2. Grouping of the SDC methods and the practice of the HCSO

There are two main types of SDC methods: (1) *perturbative* and (2) *non-perturbative* methods.

Perturbative methods are for instance rounding, adding random noise, numerical rank swapping while non-perturbative methods are global recoding, top and bottom coding, micro aggregation and cell suppression. According to the current practice of the HCSO, tabular output files are ensured to be safe by means of local *cell suppression*. It might mean notable information loss in a quantitative sense compared to the perturbative methods, but the released output will be unbiased[9].

If many values are to be suppressed, the researcher has the opportunity to restructure the tables (e.g. aggregating different categories into a less detailed one). In most cases there are more solutions (combinations of non-perturbative methods) available to regenerate a non-confidential table. The researcher is responsible for choosing that version of table which fits

[9]They do not contain rounding error or random noise so they can be regarded as „real" data.

their expectations to the highest extent. (Based on experience of checking tabular output generated in the Safe Centre it is not recommended to cross a detailed geographical breakdown with economic activity at 4-digit level. E.g. regarding a crosstab with dimensions of county and 4-digit level of NACE. In that case it is suggested to study regions instead of county or change 4-digit level to 2-digit level of NACE.)

If the SDC expert evaluates that the extent of information loss is significant in the course of output checking, a feedback is sent to the researcher. The researcher has the freedom to restructure the respective tables as the output is not released. It is important to highlight while creating the output and choosing/modifying the proper research methodology are the researcher's responsibility, choosing the proper SDC methods and making a confidentiality review about the output are the duty of the SDC expert.

The application of the procedure described above is also justified if different versions of output (e.g. crosstabs) are created (tables with different number of dimensions) from the same microdata. In that case there is a risk that while the tables could be released individually, due to mutual correspondences among the tables they become unsafe if they are released together.

> Let's consider the following example: It is assumed there are three corresponding tables with the following dimensions: *sex-town*, *sex-criminal record*, *town-criminal record*. The cell values correspond to the number of book sellers. Actually there is a three dimension table which comprised of three two dimension tables with subtotals.

Number of booksellers

Table1

	Miskolc	**Debrecen**	*Total*
Male	21	12	33
Female	16	19	35
Total	37	31	68

Table2

(Has a criminal record)	**Yes**	**No**	*Total*
Male	23	10	33

Female	8	27	35
Total	31	37	68

Table3 (Has a criminal record)	**Yes**	**No**	*Total*
Miskolc	11	26	37
Debrecen	20	11	31
Total	31	37	68

Let's consider the following notations:

— Sex: *f = male*

n = female

— Town: *M = Miskolc*

D = Debrecen

— Has a criminal record: *i = yes*

n = no

For instance X_{fDn} means the number of book sellers who are working in Debrecen and have no criminal record. This is a cell value of the three dimension table, indeed.

Based on the three dimension table the following equations can be satisfied:

- Table1

$$X_{fMi} + X_{fMn} = 21$$
$$X_{fDi} + X_{fDn} = 12$$
$$X_{nMi} + X_{nMn} = 16$$
$$X_{nDi} + X_{nDn} = 19$$

- Table2

$$X_{fMi} + X_{fDi} = 23$$
$$X_{nMi} + X_{nDi} = 8$$
$$X_{fMn} + X_{fDn} = 10$$
$$X_{nMn} + X_{nDn} = 27$$

- Table3

$$X_{fMi} + X_{nMi} = 11$$

$$X_{fMn} + X_{nMn} = 26$$

$$X_{fDi} + X_{nDi} = 20$$

$$X_{fDn} + X_{nDn} = 11$$

Solving the equations we get the following results:

$X_{fMi} = 11$ $\quad$ $X_{fMn} = 10$

$\boldsymbol{X_{fDi} = 12}$ $\quad$ $\boldsymbol{X_{fDn} = 0}$

$X_{nMi} = 0$ $\quad$ $X_{nMn} = 16$

$X_{nDi} = 8$ $\quad$ $X_{nDn} = 11$

It can be find out easily that **every male bookseller in Debrecen has a criminal record**!

3.3 Tabular data[10] protection

As it was mentioned in section 3.2, output checking can be executed via cell suppression for tabular output, which is one of the most frequently used SDC method in the case of protection of tabular data.

Hereunder the definition of cell suppression will be defined, and its usage will be illustrated on frequency and magnitude tables.

Cell suppression: a tabular data protection method, which means that some cells' value are not shown in the table but replaced by a conventional sign. If cell suppression is used for protection, we have to clearly sign which cells were suppressed (in order not to mix them with missing values). The primary and secondary suppressed cells will not be differentiated from each other, these suppressed cells will be signed with the same conventional sign.

The process of cell suppression consists of two interactive steps, the primary and secondary cell suppression, whose joint application will result in a safe tabular data. After primary cell suppression, we have to examine whether the application of secondary cell suppression is necessary, because the primary cell suppression is not enough if the value of a primarily suppressed cell can be recalculated using other cells.

Primary cell suppression: statistical disclosure control method applied to tabular data with the aim of not to disseminate but to replace with an agreed symbol (such as "…")

[10] Tabular data: data compiled into a tabular format containing aggregated information. Two essential types exist for tabular data: frequency and magnitude tables.

cells marked as sensitive based on Statistical Disclosure Control methods applied to tabular data.

Secondary cell suppression: statistical disclosure control method applied to tabular data when additional cells apart from the ones treated by primary cell suppression are suppressed in order to ensure the protection of the concerned tabular data. Its only purpose is to ensure that the value of the primarily suppressed cells remain safe. The value of the secondary suppressed cells should not be protected by itself.

The fundamental idea of the secondary cell suppression is that when values and their totals are presented in a dataset together, we want to avoid that someone is able to calculate the value of the primary suppressed cell using other cells and a total which contains it.

For example the value of the primary suppressed cell can be recalculated with basic mathematical operations from its row and column totals and from the other elements located in its row and column.

The secondary cell suppression is an optimization task, for which there are several different algorithms. There is no standard standpoint as to which suppression algorithm has to be used, but two aspects must be taken into consideration anyway:

1. Sensitive cells in tables have to be protected (values of the primarily protected cells must not be recalculated using other cells or totals.)
2. Information loss has to be as low as possible. This means that in practice:
    - We suppress as few cells as possible.
    - If there are more possible ways to suppress the fewest cells, then that solution is chosen, where the number of the contributions to the cells is the fewest.
    - Suppressing totals should be avoided.

Two essential types exist for tabular data: frequency (1) and magnitude tables (2). These two kinds of table connect to two different types of disclosure. While the frequency table makes possible only the identification of an individual, in the case of magnitude table new information can be disclosed as well. Neither identifying an individual nor releasing new information on them are permitted.

3.3.1. Protection of frequency tables

An often used alternative protection method on tabular data is rounding. The advantage of cell suppression against rounding is that the column and row totals as well as each inner, non-suppressed cell value of the table remain the same after suppression.

Number of corporations by NACE codes in 2010 (fictive values)							
Regions	**The following NACE codes were examined**						
	K	**C**	**D**	**E**	**J**	**M**	**Total**
Central-Hungary	18	500	20	11	326	1281	**2156**
Central-Transdanubia	2	145	2	15	22	140	**326**
Western- Transdanubia	4	105	5	7	10	105	**236**
Southern- Transdanubia	4	58	4	6	21	136	**229**
Northern - Hungary	1	100	*3*	5	21	119	**249**
Northern-Hungarian Plain	1	99	*8*	8	16	158	**290**
Southern- Hungarian Plain	8	107	9	8	31	163	**326**
Total	**38**	**1114**	**51**	**60**	**447**	**2102**	**3812**

*Table 1. **Example for cell suppression***

There are two basic steps of cell suppression:

First of all, the cells which need primary cell suppression are determined/marked (based on the threshold rule, generally those cells are proposed for primary suppression according to a legal act, whose value represent less than three respondents). In many cases, suppressing only the primarily marked cells (underscored cells with grey background) is not enough, because the value of these primarily suppressed cells can be calculated from row and column totals.

Due to the fact that the primary suppressed cells can be recalculated using the unsuppressed cells, other cells have to be suppressed as well. This is called secondary cell suppression. (Italic cells with dark grey background).

For example: The primary suppressed cells are:

K- Central-Transdanubia
K- Northern - Hungary
K- Northern-Hungarian Plain
D- Central-Transdanubia

The value of "K-Northern - Hungary", "K-Northern-Hungarian Plain" and "D-Central-Transdanubia" cells can be easily calculated using the row and column totals. Take into consideration that in column "*D"* secondary cell suppression has to be used because of cell "D-Central-Transdanubia, to protect both "K-Northern - Hungary" and "K-Northern-Hungarian Plain ", the suitable cells for secondary cell suppression will be selected from column "D" as well to have an optimal solution. Thus the secondary suppressed cells are "D-Northern - Hungary" and "D-Northern-Hungarian Plain ". Consequently none of the primary suppressed cells can be recalculated from others.

3.3.2. Protection of magnitude tables

In the case of magnitude tables, not only "threshold rule of three" is used, but dominance rules are also taken into consideration. The threshold rule requires the creation of the frequency table belonging to the magnitude table, while to apply the dominance rule the table which shows the share of the largest contributing firm from the cell total has to be produced (Table 4).

These tables have to be collected in the **„For_check"** folder by the researcher based on Chapter 2.

The following table is transmitted for output checking by the researcher:

Sum of corporations' income by region and examined NACE codes in 2010 (fictive values)					
Regions	The following NACE codes were examined				
	A	B	C	D	Total
1	7 767 971328.0	211 091 899 472.0	9 943 678 279.0	303 314 418 304.0	**532 117 967 383.0**
2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**
3	109 496 225 408.0	8 703 476 032.0	125 583 012 384.0	251 129 981 347.5	**494 912 695 171.5**
4	199 566 570 752.0	327 763 841 000.0	97 802 072 160.0	12 144 741 376.0	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	233 459 129 720.0	156 755 016 340.0	8 416 344 064.0	70 604 533 024.0	**469 235 023 148.0**

Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

Table 1. [11] *Example for magnitude table*

1. **The researcher produces the absolute frequency table** belonging to Table 2:
   The HCSO would like to point the researcher's attention to that in the case of calculating frequency table, the number of different respondents have to be taken into consideration.
   Assume that the threshold rule of three is used. Therefore the cell values which are smaller than 3, will be deleted (underscored, light grey cells represent primary cell suppression)

Number of corporations by NACE codes in 2010 (fictive values)					
Regions	**The following NACE codes were examined**				
	A	**B**	**C**	**D**	**Total**
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
Total	233	288	204	275	1000

Table 2. [12] *Absolute frequency table which belongs to table 2.*

2. **For each cell in Table. 2, the researcher calculates the share of the largest contribution from the cell total, and creates a new table from these shares.**
   Assume that k=90%, thus the dark grey (bold) cells will be primarily suppressed.

Share of the largest contribution from the cell total (%)					
Regions	**The following NACE codes were examined**				
	A	**B**	**C**	**D**	**Total**
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

[11] This file is put into the '*Research_output*' folder by the researcher.
[12] This file is put into the'*For_check*' folder by the researcher.

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
Total	0.77	0.72	1.05	0.71	0.20

Table 3.[13] ***Share of the largest contribution from the cell total which belongs to the Table 2.***

Primary suppressed cells are presented in Table 5. (based on the threshold rule of three and dominance rule):

Sum of the corporations' income by region and examined NACE codes in 2010 (fictive values)					
Regions	**The following NACE codes were examined**				
	A	**B**	**C**	**D**	**Total**
1	…	211 091 899 472.0	…	303 314 418 304.0	**532 117 967 383.0**
2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**
3	109 496 225 408.0	…	125 583 012 384.0	251 129 981 347.5	**494 912 695 171.5**
4	199 566 570 752.0	327 763 841 000.0	97 802 072 160.0	…	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	233 459 129 720.0	156 755 016 340.0	…	70 604 533 024.0	**469 235 023 148.0**
Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

Table 4. Table containing primary suppressed cells

It can be seen in the table below that not all suppressed cells are safe, because for example the value of the "A1", the cell value can be recalculated. In order to ensure the safety of the primary suppressed cells (marked with "…") we have to apply secondary cell suppression.

The safe output can be seen in Table 6. (after primary and secondary cell suppression)

Sum of the corporations' income by region and examined NACE codes in 2010 (fictive values)					
Regions	**The following NACE codes were examined**				
	A	**B**	**C**	**D**	**Total**
1	…	211 091 899 472.0	…	303 314 418 304.0	**532 117 967 383.0**
2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**

[13] This file is put into the '*For_check*' folder by the researcher.

3	109 496 225 408.0	**…**	125 583 012 384.0	…	**494 912 695 171.5**
4	199 566 570 752.0	…	97 802 072 160.0	**…**	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	…	156 755 016 340.0	**…**	70 604 533 024.0	**469 235 023 148.0**
Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

Table 5. Containing primary and secondary suppressed cells

Remark: In the case when row and column totals are not known, and there is a low possibility to find any other database which would contain these totals, applying secondary cell suppression is not justified.

If a magnitude table contains more than one variable (for example: income, added value, profit, etc.) then the frequency and 'share of the largest contribution from the cell total' tables have to be created for each variable separately, unless we have the same frequency and shares for each variable.

During the calculation of the frequency table, the structure of the magnitude table always has to be taken into consideration: for example if the magnitude values were calculated as the income of firms on 2 digit NACE codes, then the frequency will be the number of different firms which contributes to these income values.

4.Other SDC methods

So far the tabular data protection methods were reviewed. Hereinafter we briefly summarize other SDC methods and rules frequently applied for outputs. It is important to be stressed that in the HCSO the *principle based model* is adopted. According to this approach, the rule of thumbs approach and "*best practices*" play only a subordinate role. In contrary, we try to consider every output individually and make a decision what sensitivity rule is the most appropriate.

The threshold rule of three could be mentioned as a single rule of thumb which has to be regarded at every output comes from a population[14]. The principle based approach is beneficial, because the expected information loss is moderate compared with the rule of thumb approach. Output files are reviewed more flexibly by taking into consideration the relevant disclosure scenarios. The principle based approach requires better understanding of the output. If there is

[14]About the samples see chapter 4.1.

an ambiguous result in the output, a personal discussion between the SDC expert and the researcher might be necessary. The main drawback of the principle based approach is that output checking is more time consuming. In addition, no predefined rules or guidelines are available for the researchers which would foster creating safe output.

In the course of output checking recommendations of the principles based approach found in the *Guidelines for the checking of output based on microdata research*[15] *(ESSNet SDC, 2009)* are considered as a reference. If there is any contradiction between our guideline and the *ESSnet* Guideline (e.g. threshold rule of three vs. threshold rule of ten), then our recommendations have the priority in HCSO practice.

4.1. Methods applied to output derived from full population surveys and from sample surveys

Output generated from datasets coming from *business* surveys have to be checked taking into account the threshold rule of three. This rule applies irrespective of the type of data collection. The reasoning is that there is more public information (balance sheet, annual report) are available about companies than about the individuals or households. In general, disclosure risk is higher for business surveys than for the household surveys.

By representative household surveys and by samples drawn from population and housing census threshold rule of three is not applied automatically because of the significant information loss. In these cases we propose a consultation with the respective data owner HCSO unit. Our confidentiality reviews about the output files are in line with the recommendations of the data owner HCSO unit.

The chance of the release of a given output decreases if the survey contains *sensitive variables* (e.g. nationality, religious affiliation, sexual orientation, health, income etc.).

During output checking, special attention is devoted to *identifying variables*, i.e.:

- geographical breakdown (e.g.: county, settlement, enumeration area etc.)
- classifications (for example. HSCO[16], CPA[17], NACE, ANIE[18], etc.);

[15]Download from: http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

[16] **H**ungarian **S**tandard **C**lassification of **O**ccupations

[17] **C**lassification of **P**roducts by **A**ctivity

- age;
- sex;
- marital status;
- highest educational attainment;
- number of children in the household;
- size of the household.

5.Advices for researchers to create safe research output

In order to create safe research output the following hints and tips are to be considered:

- Please submit only the final research output for output checking, if possible.
- Please consider what are the most important and most relevant research outputs which you really need:
    - If the research output contains too many files or they are too big, output checking will last longer. The SDC expert has to understand the goal and the main steps of the research project and assess their coherence at a certain level as well.
    - It could pose a risk of an output will be categorized as unsafe because of correspondence among the tables, if many files are intended to be released or there are many variables in the tables.
- Do you observe low number of respondents (n=1,2) in the frequency count tables or in the magnitude tables? Keep in your mind that tabular data output files are always checked whether they violate the threshold rule of three or not.
- If the proportion of cells with the low number respondents' is high enough, consider restructuring the table by means of dimension reduction, aggregation etc. The cells marked as sensitive are suppressed.
- Are there row- or column totals in the tables? If yes, implementation of secondary cell suppression could be also necessary which could result in notably higher number of suppressed cells.

[18] **A**ctivity **N**omenclature of **I**ndividual **E**ntrepreneurs

- Writing a code (script) is preferred in the course of preparing output files, because the results will be more traceable and reproducible.
- If an output was generated by means of a code, create the corresponding log file derives from the entire run as well. Copy the log file into the *'For_ check'* folder. Log files are very helpful to understand results quicker and better.
- We recommend you produce the entire output with one single code, because the procedure of the research will be more transparent. If you rather prefer to write several pieces of code please consider the following:
    - Use a "master" code in which you make references to the other dependent codes.
    - Create a separate text file in which you denote the logical order of the codes. Copy this file into the *'For_check'* folder.
- If you save a dataset several times during the research please do not overwrite the previous versions of them, because your results become irreproducible. If only the latest version of the dataset is available, calculations based on previous versions will be impossible to perform.
- Save the code before you run it. Do not run codes from the 'Temp' folder with .tmp extension, because the code to run cannot be specified.

It happens on a regular basis that preparing safe output files need recalibration and redesign of the models, because only modifying the results is not satisfactory. The principle of researchers' freedom implies that solely the researcher takes the responsibility for the professional content of their results. So the researcher is entitled and supposed to choose the proper alternative methods.

Documentation Form[1] of the Research Output

1. Basic information about the research

1.1 Researcher's username:	
1.2 Name of the research project:	
1.3 Name of the Institution:	
1.4 Name of the researcher producing the output:	
1.5 Researcher's e-mail address:	
1.6 Date of the output request (YYYY/MM/DD):	

2. Documentation for the research output

2.1 Research output produced by program scripts[2]

Names of the research output files[3]	Type of the research output files[4]	Please give the name(s) and line(s) of program script(s) that produce the research output	Please give the name of the original survey(s)/ input data set(s)[5], from which the research output is produced

[1] Output checking can only start after the researcher has completely filled out the form, and submitted it to the dissemination unit.

[2] Program scripts that produce the research output are available. (for example: STATA, SAS scripts).

[3] Each file can be listed in one cell if they were produced from the same microdata and were generated from the same row of the script.

[4] For example frequency table, magnitude table, regression, log file. If the research output is a log file, then results (regressions, aggregated tables) do not have to be listed, but they have to be commented in the scripts.

[5] For example. HCSO-input file: Census, Mortality, R&D data sets etc.; Input which was brought by the researcher with its exact name.

2.2 Research output produced without [6] program scripts

2.2.1 Please briefly describe the main purpose, logical process of the research output and its main steps.

2.2.2 Please give the definition of the new or renamed variables which can be found in the output or in the in the datasets used to create the research output[7]!

[6] In this case no scripts were used to produce the research output. (for example: Excel, SPSS).

[7] If the variables have already been defined in a separate document (for example: World, Excel) then this table does not have to be filled out. Please put this document into the 'Documentation_form' folder.

Name of the variable	Definition of the variable	Which dataset contains it

2.2.3 Please describe the research output[8]!

Name of the research output file	Type of the research output[9] file	Which dataset contains it	Brief description of the result

[8] If the description of the research output has already been presented in an output file, then indicating the output file name is enough (for example <x_y><xyz> file).

[9] For example frequency table, magnitude table, regression, log file. If the research output is a log file (for example .spv output), then all results have to be listed and described.

The research output can only be released from HCSO's safe environment after the Documentation Form of the Research Output, examined by the staff of the dissemination unit, has been properly filled in.

The research output will be forwarded to the expert for output checking only if the Documentation Form is filled in completely and properly.

The dissemination unit will give feedback to the researcher about the correctness of the Documentation Form within 2-3 days.

If the Documentation Form is filled in incorrectly or incompletely then the researcher will be asked to correct/complete the documentation.

Each field has to be filled in taking into consideration the following remarks:

If each research output is generated by scripts, then it is enough to fill in part 2.1, otherwise part 2.2 also has to be completed.