REPUBLIKA SLOVENIJA
STATISTIČNI URAD RS

Getting started with Mu Argus

SURS

Input files for Mu Argus

Basic terms in Mu Argus Variable types

- **HH Identifier:** The unique identifier of a household
- **HH Variable:** A variable that by its nature has the same value for each member of a household
- **Weight:** The variable is a sampling weight
- **Categorical:** Can be defined as a quasi-identifier
- **Numerical:** A numerical variable can be used for top/bottom coding, microaggregation and rounding

- A variable can be both numerical and categorical (=ordinal)

Basic terms in Mu-Argus

- **The weight for local suppression**, default value is 50. A higher value means less possibility for suppression.
- The name of a **codelist file is optional** (it is only used when displaying information on this variable).
- It is also possible to specify the **truncation** if it is a feasible way of recoding (special case of hierarchical variable)
- **At least one missing value has to be specified for each categorical variable**. Missing values play a specific role in the SDC-process, as missing values will be imputed when local suppression is applied.
  - – The weight variable cannot have a missing value.

Truncation

- In case of hierarchical structure of variable‘s codes
- Certain number of characters is chopped from the end of variable‘s values

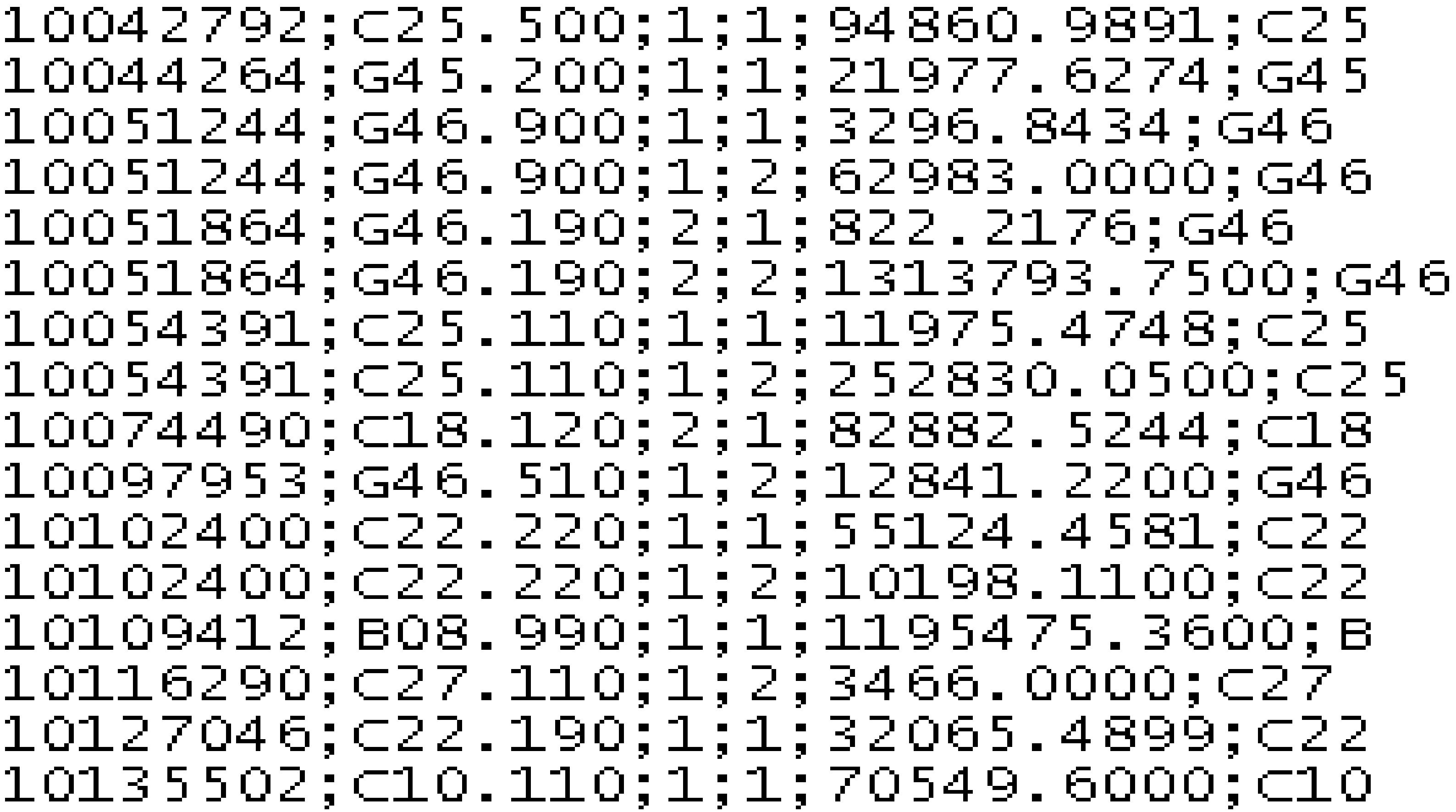
- Example:
  - A10.100 (NACE)
  - 4 characters are chopped from NACE
  - Result: A10

Input files with microdata

- Only the variables that are the quasi-identifiers need to be in the input file for Mu-Argus.
  - Statistical identifier should be also included in the input file, after protection all non-confidential variables are added (link is statistical identifier).
- Members of the same households must be grouped together (sorted by the household identifier).
- Sampling weights must not have missing values.
- Structure
  a. a fixed format ASCII file (.asc)
  b. free format file with a specified separator (.csv) with or without variable names
  c. SPSS format

Microdata file (.csv)

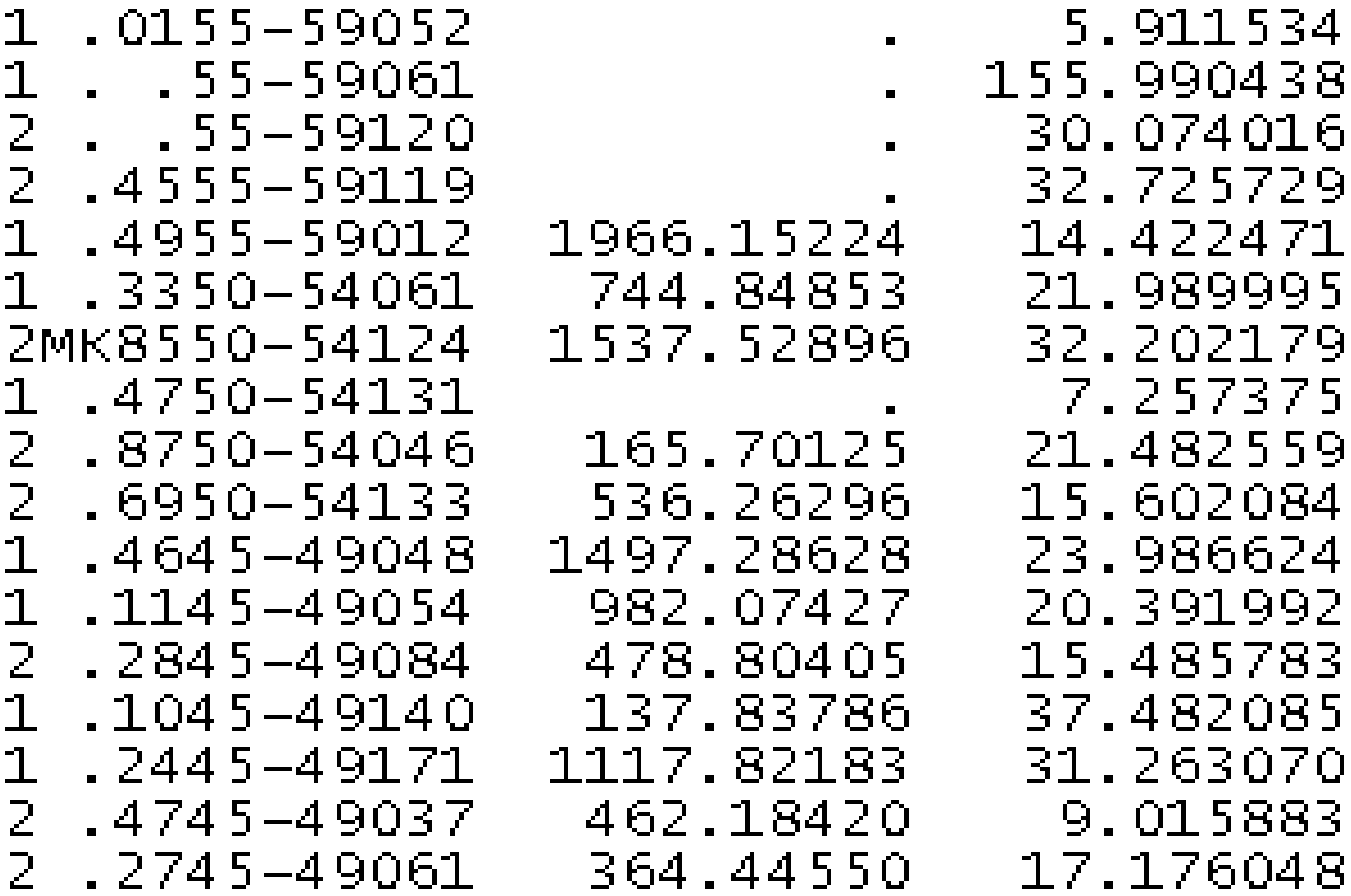
- Delimited file with or without variable names in the first row



```
10042792;C25.500;1;1;94860.9891;C25
10044264;G45.200;1;1;21977.6274;G45
10051244;G46.900;1;1;3296.8434;G46
10051244;G46.900;1;2;62983.0000;G46
10051864;G46.190;2;1;822.2176;G46
10051864;G46.190;2;2;1313793.7500;G46
10054391;C25.110;1;1;11975.4748;C25
10054391;C25.110;1;2;252830.0500;C25
10074490;C18.120;2;1;82882.5244;C18
10097953;G46.510;1;2;12841.2200;G46
10102400;C22.220;1;1;55124.4581;C22
10102400;C22.220;1;2;10198.1100;C22
10109412;B08.990;1;1;1195475.3600;B
10116290;C27.110;1;2;3466.0000;C27
10127046;C22.190;1;1;32065.4899;C22
10135502;C10.110;1;1;70549.6000;C10
```

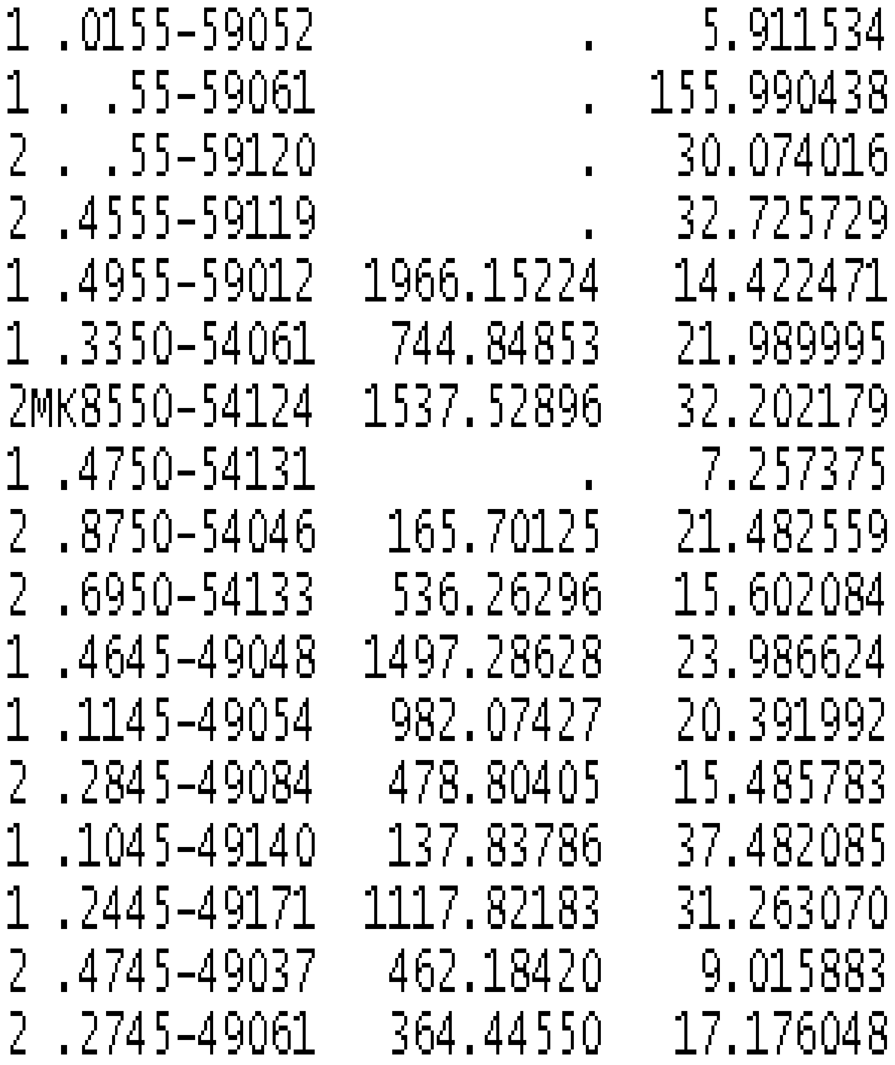
Microdata file (.asc)

- Variables have fixed length



```
1 .0155-59052           .    5.911534
1 . .55-59061           .  155.990438
2 . .55-59120           .   30.074016
2 .4555-59119           .   32.725729
1 .4955-59012  1966.15224   14.422471
1 .3350-54061   744.84853   21.989995
2MK8550-54124  1537.52896   32.202179
1 .4750-54131           .    7.257375
2 .8750-54046   165.70125   21.482559
2 .6950-54133   536.26296   15.602084
1 .4645-49048  1497.28628   23.986624
1 .1145-49054   982.07427   20.391992
2 .2845-49084   478.80405   15.485783
1 .1045-49140   137.83786   37.482085
1 .2445-49171  1117.82183   31.263070
2 .4745-49037   462.18420    9.015883
2 .2745-49061   364.44550   17.176048
```

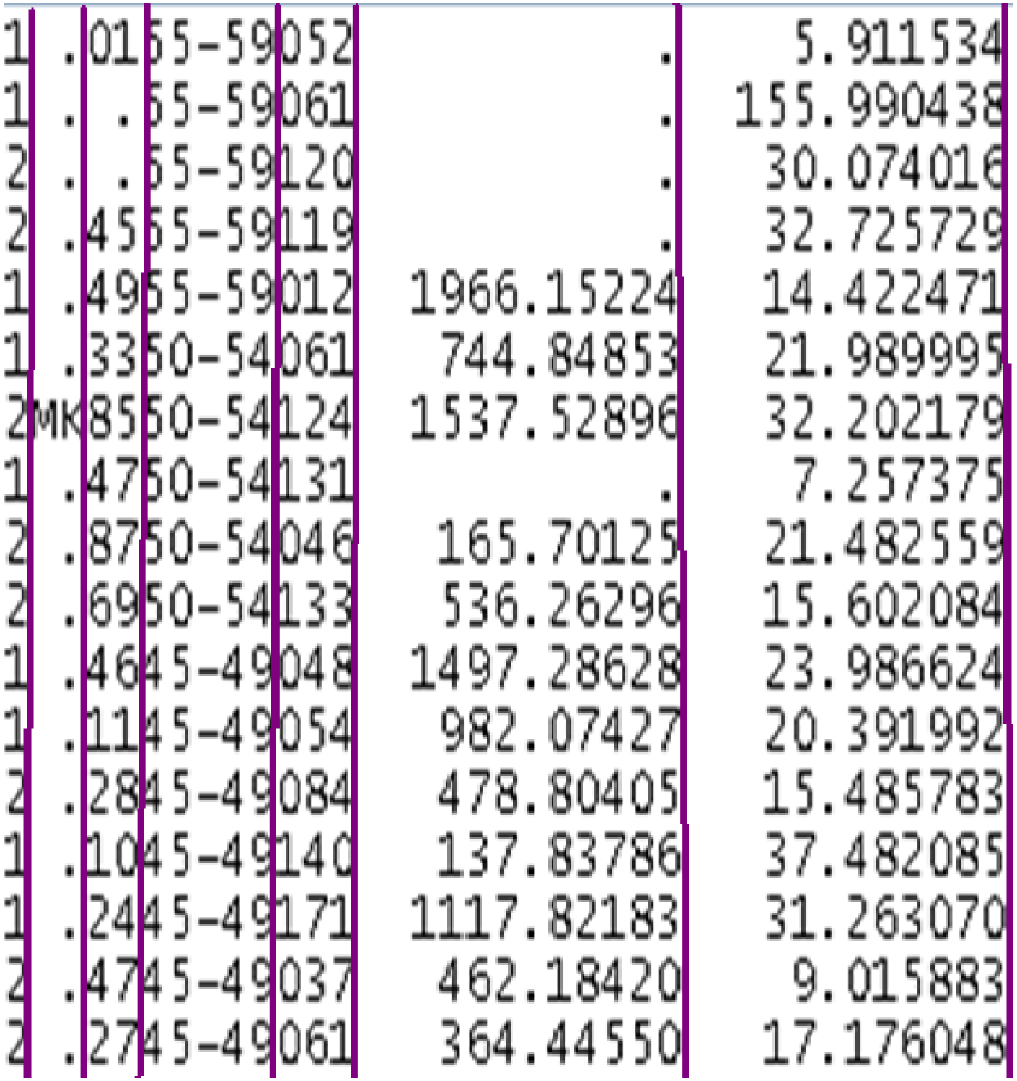
Structure of ASCII file



```
1 .0155-59052           .    5.911534
1 . .55-59061           .  155.990438
2 . .55-59120           .   30.074016
2 .4555-59119           .   32.725729
1 .4955-59012  1966.15224   14.422471
1 .3350-54061   744.84853   21.989995
2MK8550-54124  1537.52896   32.202179
1 .4750-54131           .    7.257375
2 .8750-54046   165.70125   21.482559
2 .6950-54133   536.26296   15.602084
1 .4645-49048  1497.28628   23.986624
1 .1145-49054   982.07427   20.391992
2 .2845-49084   478.80405   15.485783
1 .1045-49140   137.83786   37.482085
1 .2445-49171  1117.82183   31.263070
2 .4745-49037   462.18420    9.015883
2 .2745-49061   364.44550   17.176048
```

- Right-aligned variables
- No variable names in the first row
- Missing values are allowed
- All values for each numerical variable have to have the same number of decimal places

Structure of ASCII file



```
1 .0155-59052           .    5.911534
1 . .55-59061           .  155.990438
2 . .55-59120           .   30.074016
2 .4555-59119           .   32.725729
1 .4955-59012  1966.15224   14.422471
1 .3350-54061   744.84853   21.989995
2MK8550-54124  1537.52896   32.202179
1 .4750-54131           .    7.257375
2 .8750-54046   165.70125   21.482559
2 .6950-54133   536.26296   15.602084
1 .4645-49048  1497.28628   23.986624
1 .1145-49054   982.07427   20.391992
2 .2845-49084   478.80405   15.485783
1 .1045-49140   137.83786   37.482085
1 .2445-49171  1117.82183   31.263070
2 .4745-49037   462.18420    9.015883
2 .2745-49061   364.44550   17.176048
```

- gender – 1 place
- citizenship – 2 places
- activity (NACE) – 2 places
- age classes– 5 places
- municipality – 3 places
- income – 12 places, 5 decimal places
- weight - 12 places, 5 decimal places
- Decimal point is 1 place long!

Metadata file (.rda)

- The description of input data file (structure of .asc or .csv)
- Includes metadata about variables
  - Identification level
  - Missing values
  - Length + number of decimals
  - Type of variable (response, household identifier, explanatory, weight; numerical/categorical, etc.)
  - Links between variables are specified (the same suppression pattern).
- Differs due to the type of input file (.asc or .csv)
- It can be created in Mu Argus.

Metadata file (.rda)

METADATA FILE	MEANING
<RECODABLE>	This variable may be recoded.
<CODELIST>	Name of the codelist file
<IDLEVEL>	Identification level
<TRUNCABLE>	Relevant way of recoding (e.g. NACE)
<NUMERIC>	The variable is numeric.
<DECIMALS>	The number of decimal positions for a (numeric) variable
<WEIGHT>	The variable contains sample weights.

Metadata file (.rda)

METADATA FILE	MEANING
<HOUSE_ID>	This variable is a household identification.
<HOUSEHOLD>	A household variable typically contains the same value for each member of a household. When the suppression of the value for one member is necessary, it will be done for all members.
<SUPPRESSWEIGHT>	Priority weight for the selection of the suppression pattern; default value = 50

Metadata file for .asc file

```
REGION 1  4 9999 9998
  <RECODABLE>
  <CODELIST> "regio.cdl"
  <IDLEVEL>  1
  <SUPPRESSWEIGHT>  50
  <TRUNCABLE>
SEX 5  1  9
  <RECODABLE>
  <CODELIST> "Sex.cdl"
  <IDLEVEL>  2
  <SUPPRESSWEIGHT>  50
MARSTAT 8  1 9
  <RECODABLE>
  <IDLEVEL>  3
  <SUPPRESSWEIGHT>  50
```

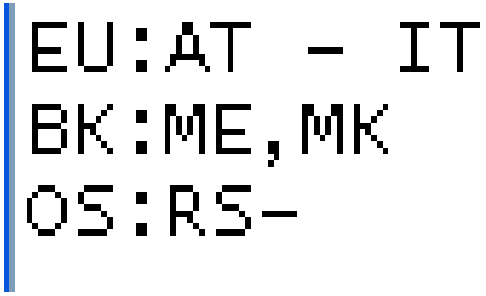
Variable‘s length

Missing value

Starting place

File for global recoding (.grc)

- Categorical variable
- We can write it in Mu-Argus and save it as .grc file or we can import .grc file.
- Structure:
  - – on the left new value, after the colon recoded values:



```
EU:AT - IT
BK:ME,MK
OS:RS-
```

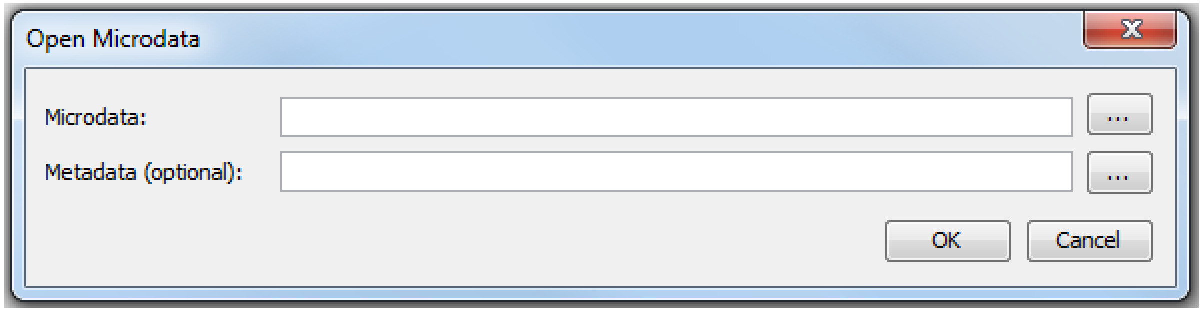
Codelist file (.cdl)

- Each categorical variable can have a code list.
- Code lists are used only in Mu Argus.

```
1,Dutch
2,North-Europe
3,South-Europe
4,North-America
5,South-America
6,Mediterrenean
7,African
8,Asian
9,Unknown
```

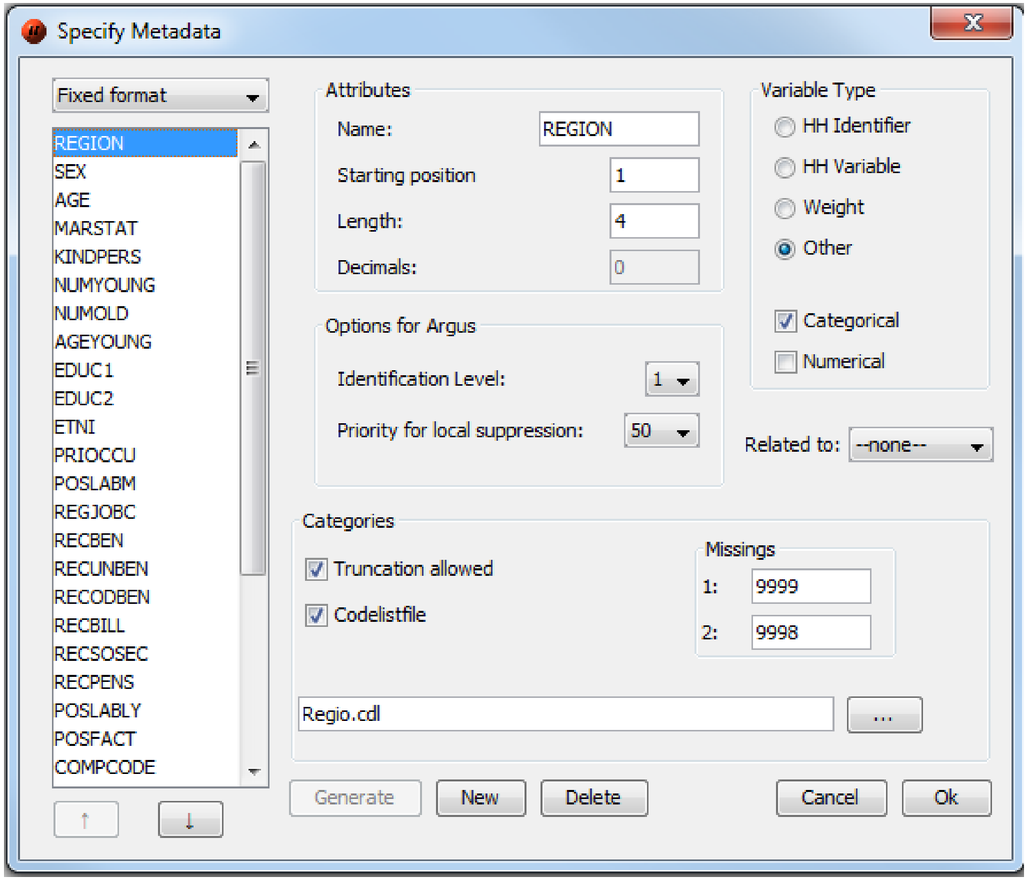
Introduction to Mu Argus

File | Open Microdata



- The menu for choosing the microdata (.asc,.csv) and optionally the metadata file („.rda“)

Specify | Metafile


Specify Metadata

Fixed format

REGION
SEX
AGE
MARSTAT
KINDPERS
NUMYOUNG
NUMOLD
AGEYOUNG
EDUC1
EDUC2
ETNI
PRIOCCU
POSLABM
REGJOBC
RECBEN
RECUNBEN
RECODBEN
RECBILL
RECSOSEC
RECPENS
POSLABLY
POSFACT
COMPCODE

Attributes

Name: REGION
Starting position 1
Length: 4
Decimals: 0

Variable Type

○ HH Identifier
○ HH Variable
○ Weight
● Other

☑ Categorical
☐ Numerical

Options for Argus

Identification Level: 1
Priority for local suppression: 50

Related to: --none--

Categories

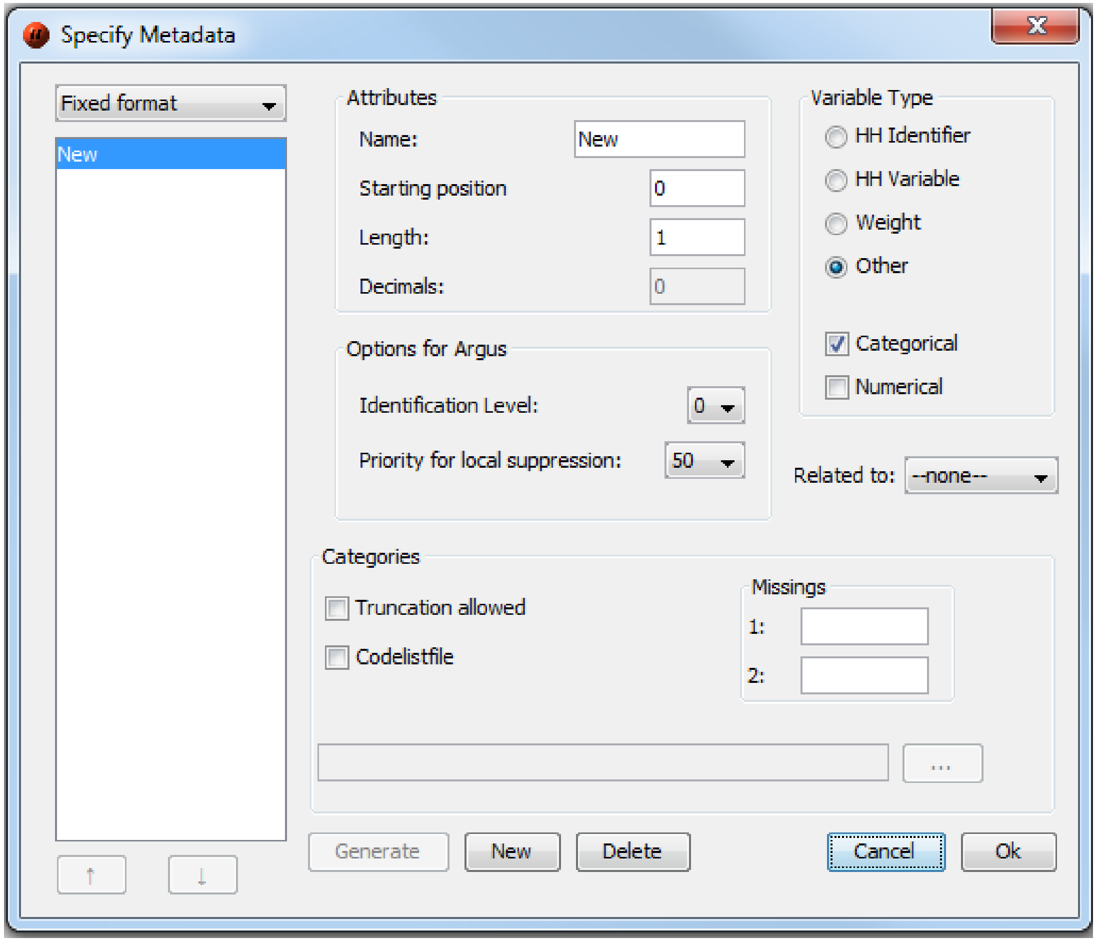
☑ Truncation allowed
☑ Codelistfile

Missings
1: 9999
2: 9998

Regio.cdl

- Construction / Change of metadata file (.rda)
- Variable properties:
  - Length, starting position, missing values, codelists, identification level, etc.
  - Related to …

Specify | Metafile



Specify Metadata

Fixed format

New

Attributes

Name: New

Starting position 0

Length: 1

Decimals: 0

Options for Argus

Identification Level: 0

Priority for local suppression: 50

Variable Type

○ HH Identifier

○ HH Variable

○ Weight

● Other

☑ Categorical

□ Numerical

Related to: --none--

Categories

□ Truncation allowed

□ Codelistfile

Missings

1:

2:

- In case of no „.rda“ file, click button „New“ to enter variables‘ metadata.
- *Fixed format* – „.asc“, *Free format* – „.csv“, *Free format with meta* – „.csv“ with variables‘ names in the first row.

Identification levels

- **0**: an individual cannot be identified by this variable and it will not play a role in the disclosure control process.
- **1**: the variable is most identifying (E)
- **2**: the variable is more identifying (V)
- **3**: the variable is identifying (I)

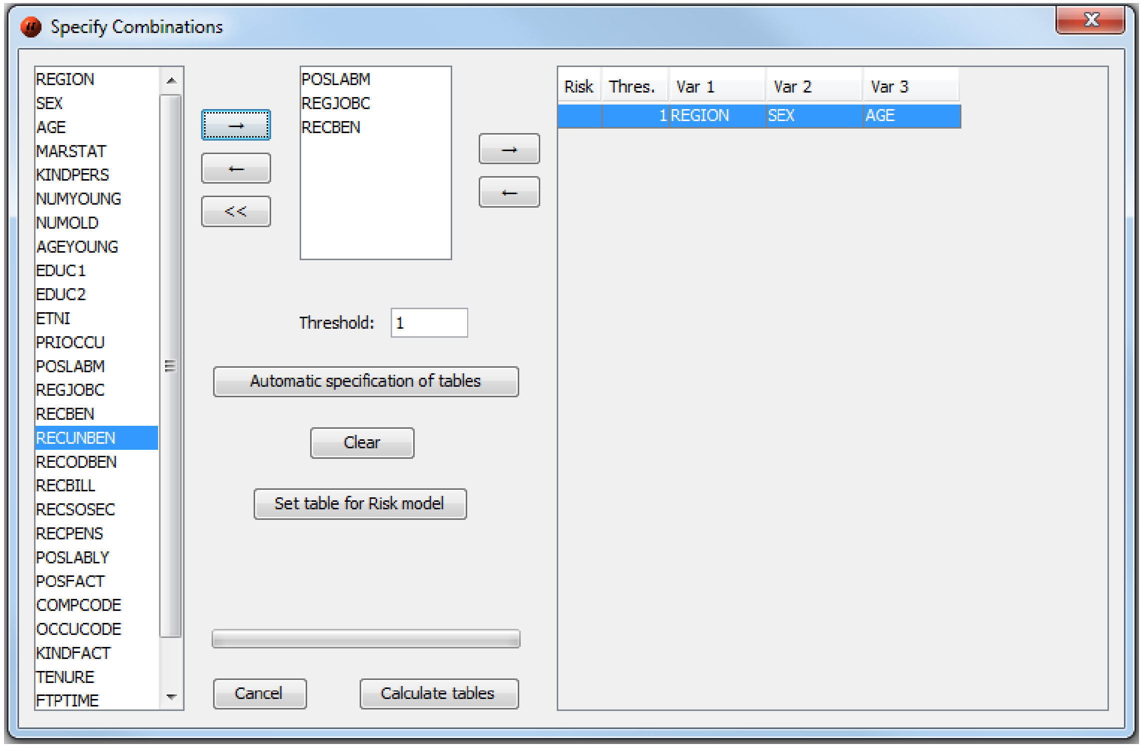
Specify | Combinations

- Manually specified
- Automatic specification of tables, identification level > 0
  - Identification levels used:

$$E \times V \times I \quad (E \leq V \leq I)$$

  - All tables up to the given dimension are calculated, for each dimension a threshold can be specified
    - Threshold is the maximum number of combinations **still considered unsafe!**
- In case of a sample, the frequencies are calculated on a sample.

Specify | Combinations



REGION
SEX
AGE
MARSTAT
KINDPERS
NUMYOUNG
NUMOLD
AGEYOUNG
EDUC1
EDUC2
ETNI
PRIOCCU
POSLABM
REGJOBC
RECBEN
RECUNBEN
RECODBEN
RECBILL
RECSOSEC
RECPENS
POSLABLY
POSFACT
COMPCODE
OCCUCODE
KINDFACT
TENURE
FTPTIME

POSLABM
REGJOBC
RECBEN

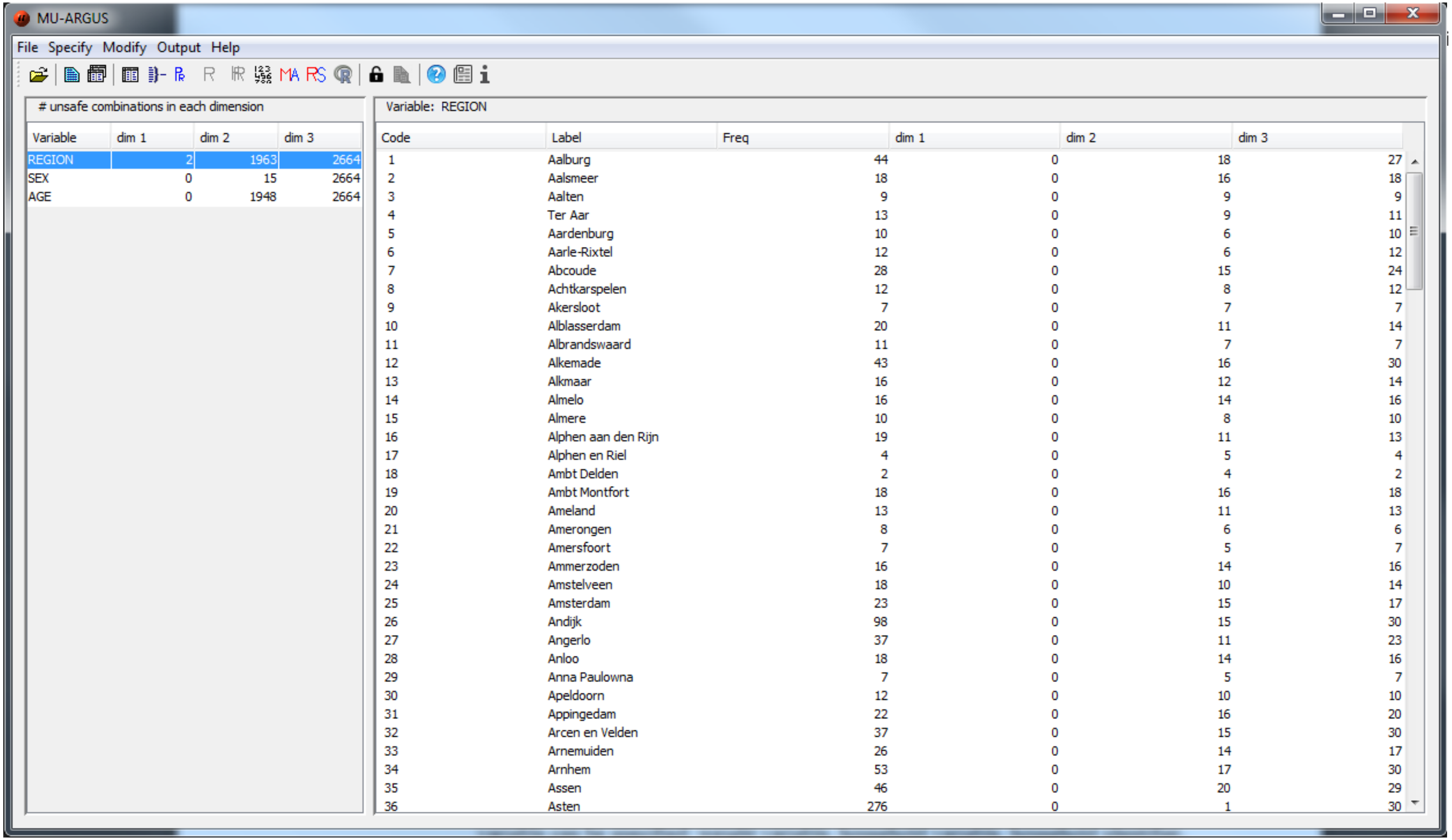
Threshold: 1

Risk	Thres.	Var 1	Var 2	Var 3
	1	REGION	SEX	AGE

Three options:

- Specified manually
- Automatic specification of tables
  - Identification levels
  - Up to given dimension
- Special combination can be selected for risk estimation

- Click Calculate tables.



unsafe combinations in each dimension

Variable	dim 1	dim 2	dim 3
REGION	2	1963	2664
SEX	0	15	2664
AGE	0	1948	2664

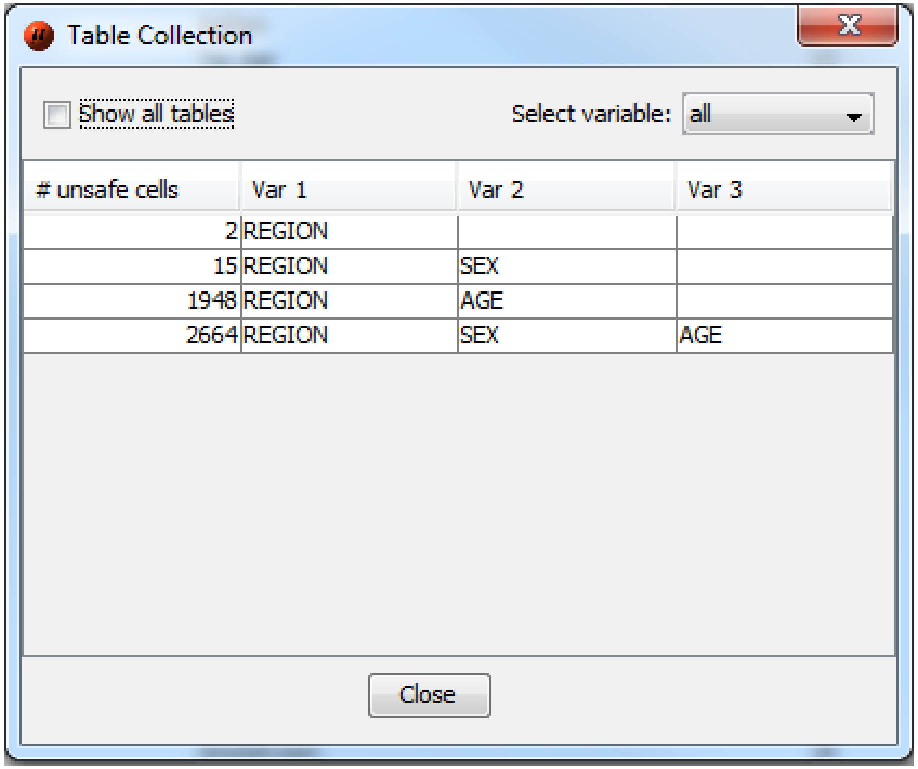
Variable: REGION

Code	Label	Freq	dim 1	dim 2	dim 3
1	Aalburg	44	0	18	27
2	Aalsmeer	18	0	16	18
3	Aalten	9	0	9	9
4	Ter Aar	13	0	9	11
5	Aardenburg	10	0	6	10
6	Aarle-Rixtel	12	0	6	12
7	Abcoude	28	0	15	24
8	Achtkarspelen	12	0	8	12
9	Akersloot	7	0	7	7
10	Alblasserdam	20	0	11	14
11	Albrandswaard	11	0	7	7
12	Alkemade	43	0	16	30
13	Alkmaar	16	0	12	14
14	Almelo	16	0	14	16
15	Almere	10	0	8	10
16	Alphen aan den Rijn	19	0	11	13
17	Alphen en Riel	4	0	5	4
18	Ambt Delden	2	0	4	2
19	Ambt Montfort	18	0	16	18
20	Ameland	13	0	11	13
21	Amerongen	8	0	6	6
22	Amersfoort	7	0	5	7
23	Ammerzoden	16	0	14	16
24	Amstelveen	18	0	10	14
25	Amsterdam	23	0	15	17
26	Andijk	98	0	15	30
27	Angerlo	37	0	11	23
28	Anloo	18	0	14	16
29	Anna Paulowna	7	0	5	7
30	Apeldoorn	12	0	10	10
31	Appingedam	22	0	16	20
32	Arcen en Velden	37	0	15	30
33	Arnemuiden	26	0	14	17
34	Arnhem	53	0	17	30
35	Assen	46	0	20	29
36	Asten	276	0	1	30

Number of unsafe combinations for each dimension/variable.

If n-dimensional combination is checked, then all i-dimensional combinations are also checked, i = 1…n-1!

Modify | Show Table Collection



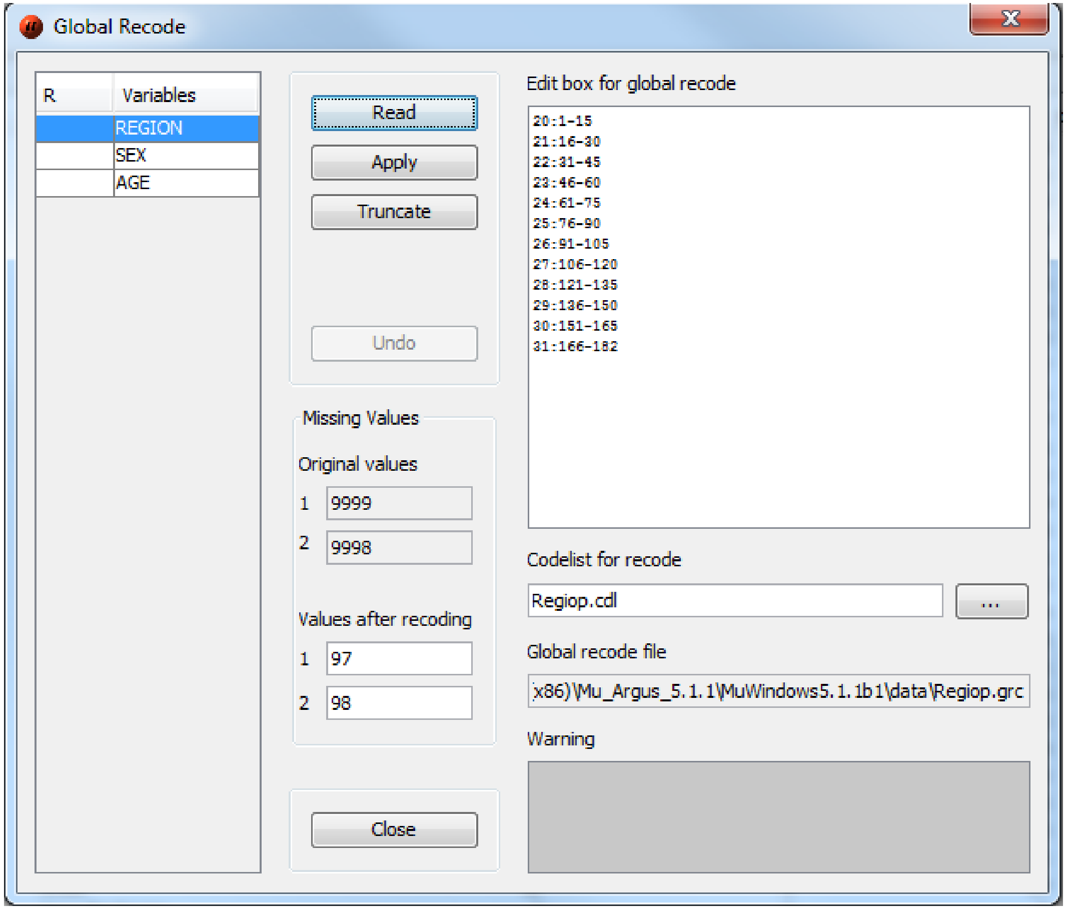
□ Show all tables

Select variable: all

# unsafe cells	Var 1	Var 2	Var 3
2	REGION		
15	REGION	SEX	
1948	REGION	AGE	
2664	REGION	SEX	AGE

- Select variable – tables with just chosen variable

Modify | Global recode



R	Variables
	REGION
	SEX
	AGE

Edit box for global recode

```
20:1-15
21:16-30
22:31-45
23:46-60
24:61-75
25:76-90
26:91-105
27:106-120
28:121-135
29:136-150
30:151-165
31:166-182
```

Missing Values

Original values

1 9999

2 9998

Values after recoding

1 97

2 98

Codelist for recode: Regiop.cdl

Global recode file: x86)\Mu_Argus_5.1.1\MuWindows5.1.1b1\data\Regiop.grc

Warning

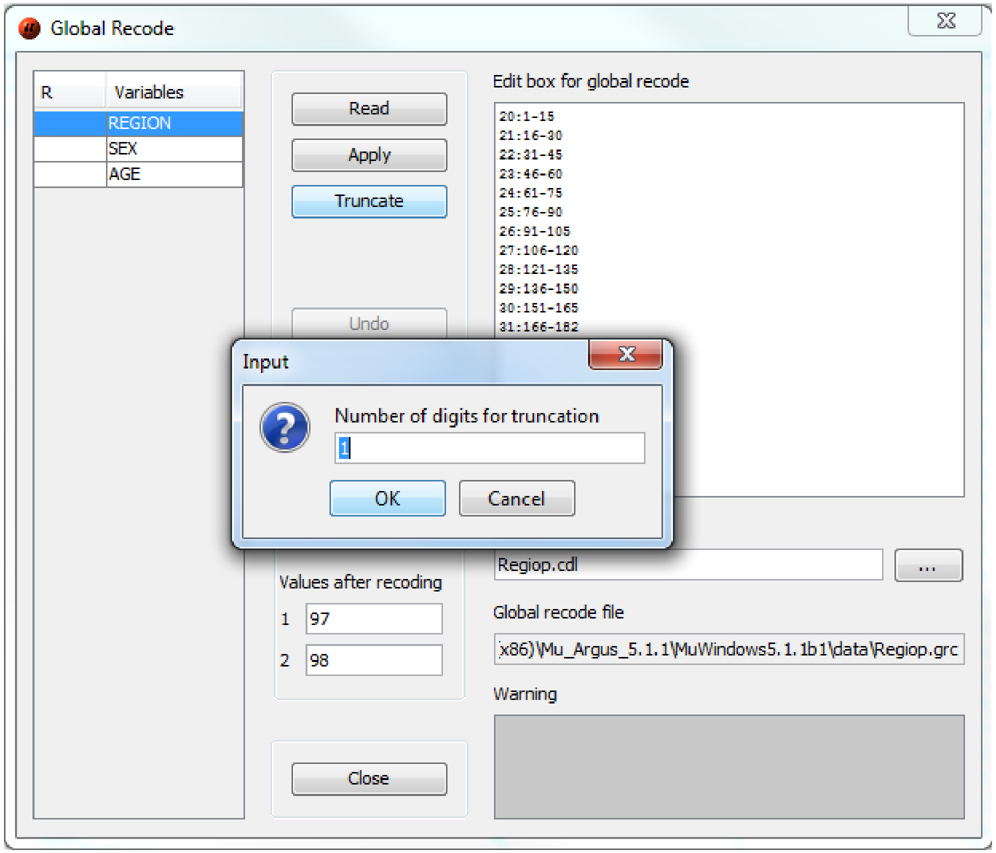
- Read – import of „.grc“ file

  OR

  Write it manually

- **Don‘t forget to click Apply!**

Modify | Global recode



R	Variables
	REGION
	SEX
	AGE

Edit box for global recode

```
20:1-15
21:16-30
22:31-45
23:46-60
24:61-75
25:76-90
26:91-105
27:106-120
28:121-135
29:136-150
30:151-165
31:166-182
```

Number of digits for truncation: 1

Regiop.cdl

Values after recoding

1 97

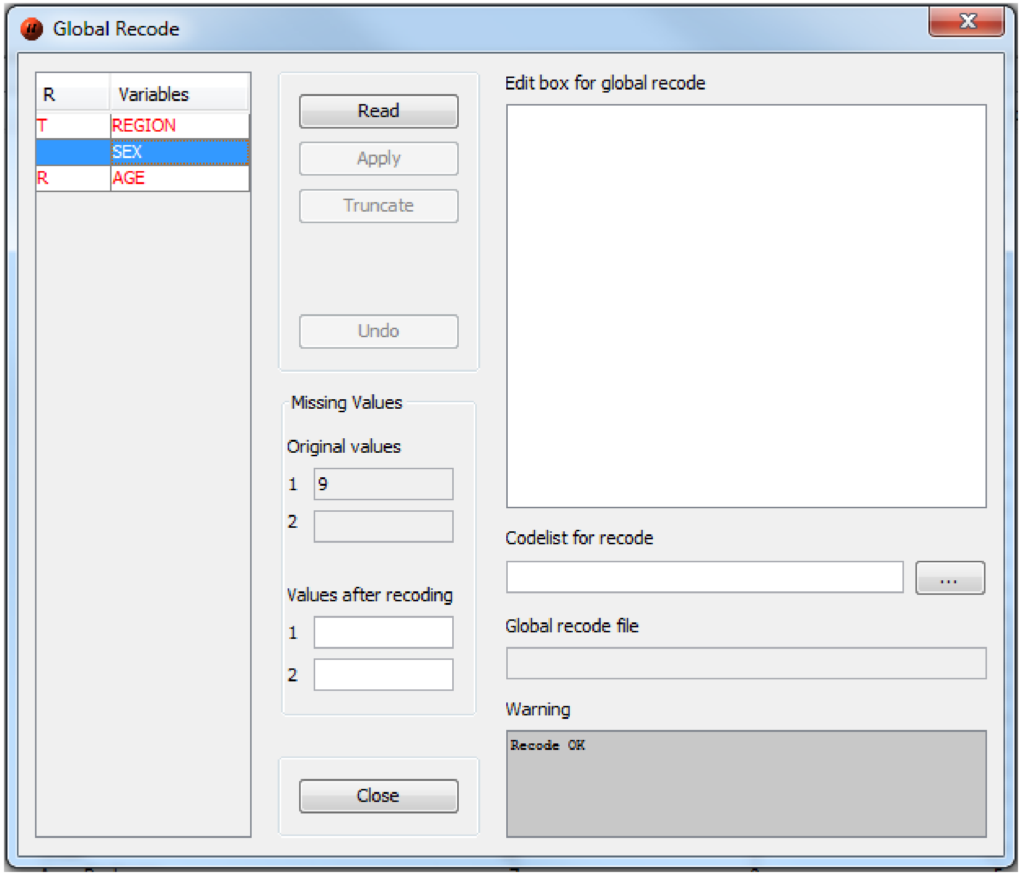
2 98

Global recode file: x86)\Mu_Argus_5.1.1\MuWindows5.1.1b1\data\Regiop.grc

Warning

- Truncate
  - Specify the number of characters
  - x characters are chopped from the end of variable‘s values (special case of hierarchical variable)
  - Always applied to the original values (if you want to truncate the same variable twice, each time one digit, you have to fill in “2” the second time)

Modify | Global recode



R	Variables
T	REGION
	SEX
R	AGE

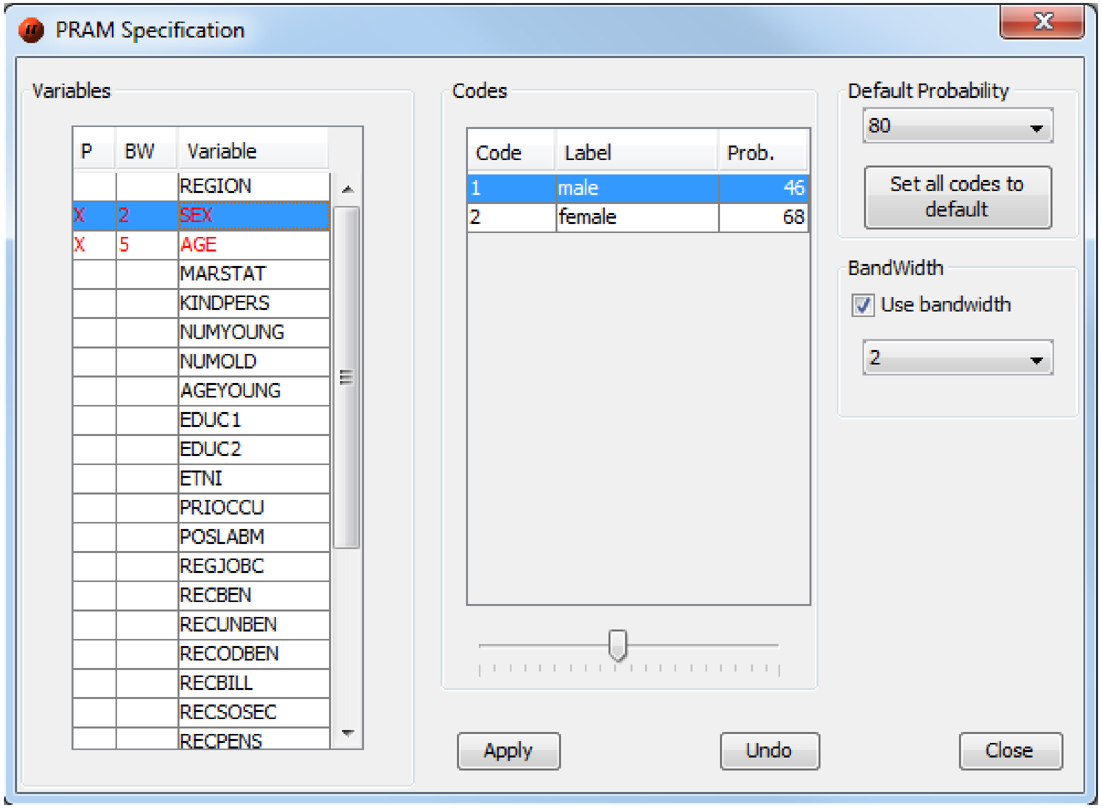
Original values
1 9
2

Values after recoding
1
2

Warning
Recode OK

- If you apply a global recoding or truncate a variable, the colour of the variable will be changed into red and an ‘R’ or a ‘T‘ will be indicated in the first column of the list-window.

Modify | PRAM specification



PRAM Specification

Variables

P	BW	Variable
		REGION
X	2	SEX
X	5	AGE
		MARSTAT
		KINDPERS
		NUMYOUNG
		NUMOLD
		AGEYOUNG
		EDUC1
		EDUC2
		ETNI
		PRIOCCU
		POSLABM
		REGJOBC
		RECBEN
		RECUNBEN
		RECODBEN
		RECBILL
		RECSOSEC
		RECPENS

Codes

Code	Label	Prob.
1	male	46
2	female	68

Default Probability: 80

BandWidth

☑ Use bandwidth

2

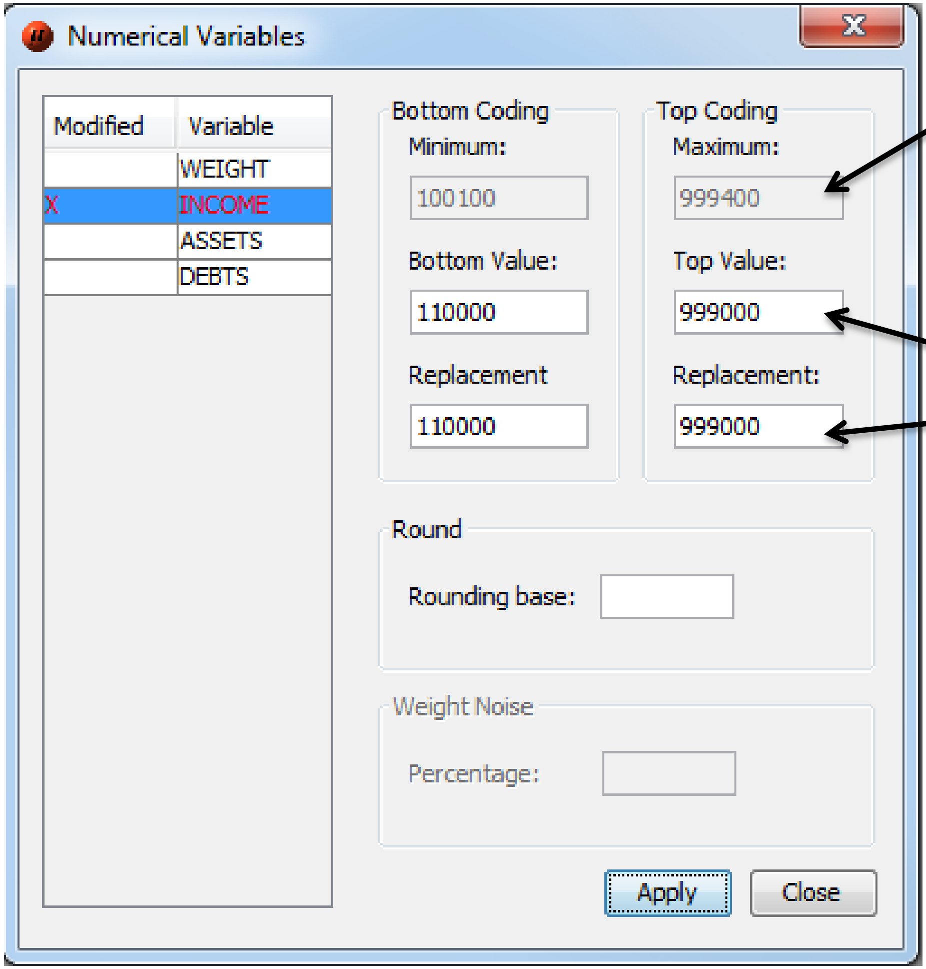
- Default probability – probability that values are not changed
- Use bandwidth – changing the value is limited to the nearest n values

- Use of PRAM is shown in the listbox by an X in the first column and an indication whether the bandwidth has been used or not.

Modify | Modify numerical variables

- Top/Bottom coding
- Rounding
- Add noise to the weight variable

Modify | Modify numerical variables | Top/Bottom coding



Numerical Variables

Modified	Variable
	WEIGHT
X	INCOME
	ASSETS
	DEBTS

Bottom Coding
Minimum: 100100
Bottom Value: 110000
Replacement 110000

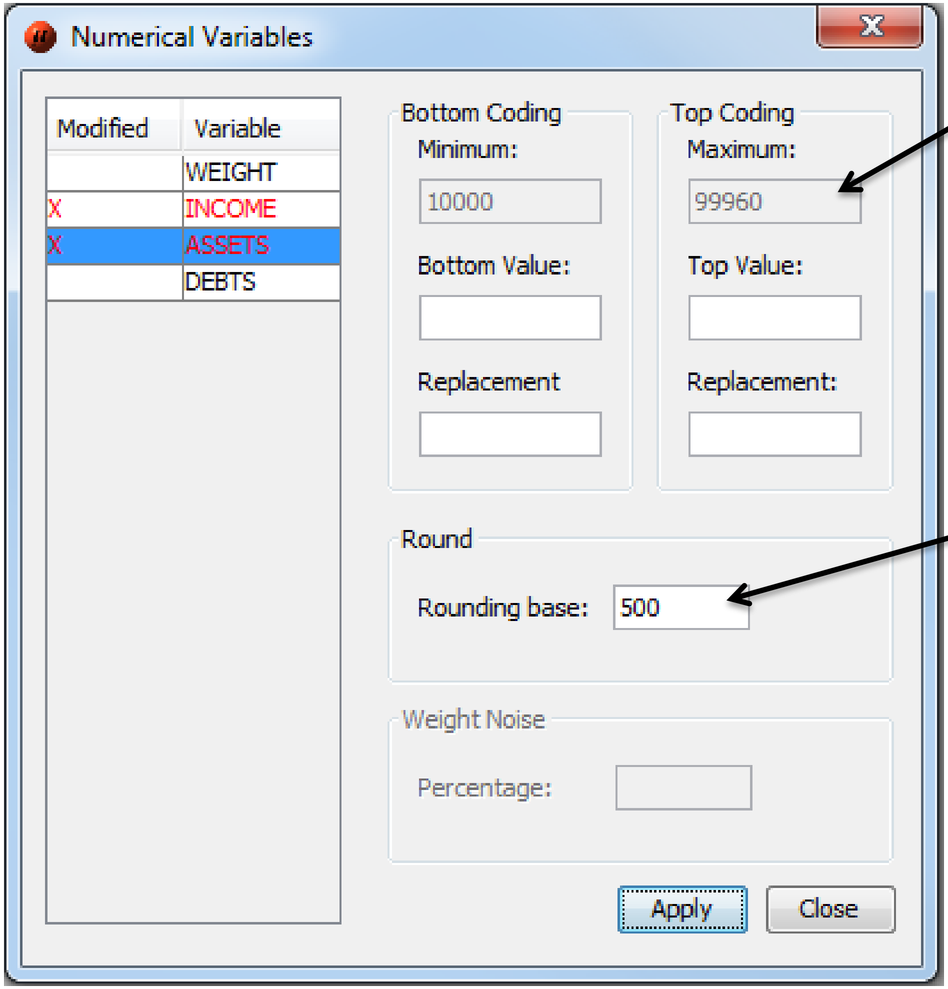
Top Coding
Maximum: 999400
Top Value: 999000
Replacement: 999000

Round
Rounding base:

Weight Noise
Percentage:

- Actual top and bottom value for variable INCOME.
- Values below/over these thresholds are replaced. **Click Apply.**
- Use of the method is shown in the listbox by an X.

Modify | Modify numerical variables | Round



Numerical Variables

Modified	Variable
	WEIGHT
X	INCOME
X	ASSETS
	DEBTS

Bottom Coding
Minimum: 10000
Bottom Value:
Replacement

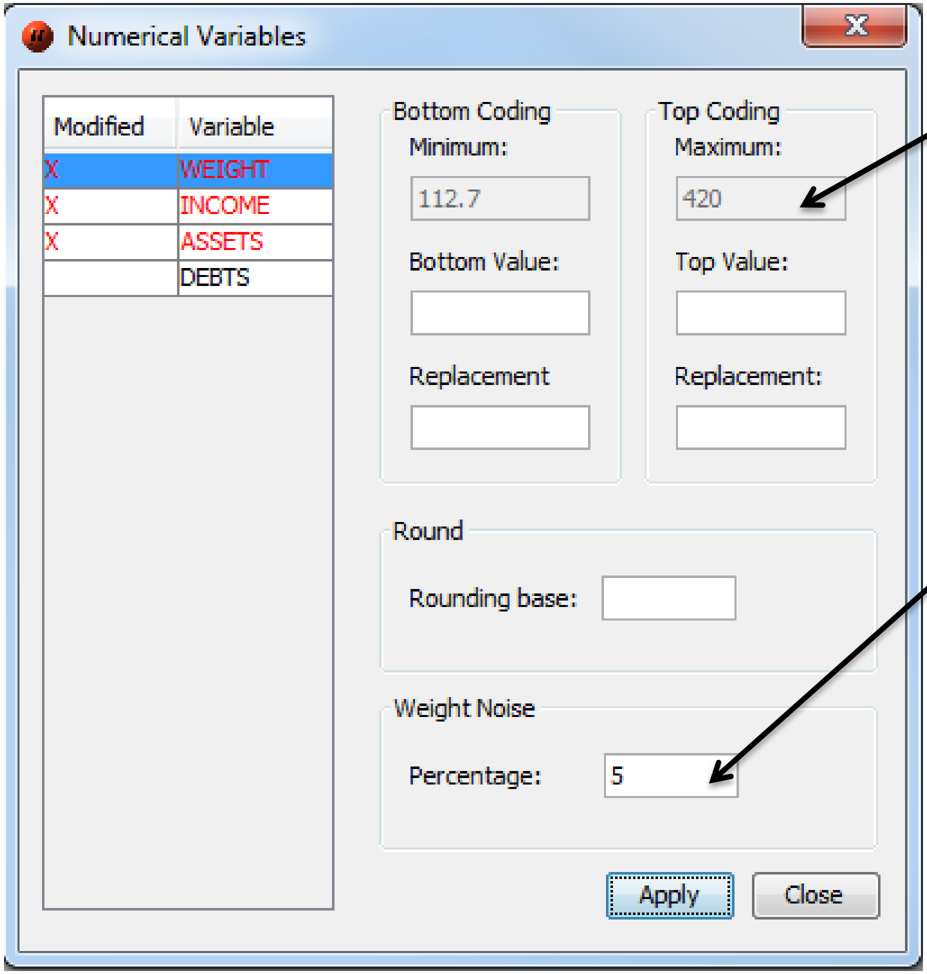
Top Coding
Maximum: 99960
Top Value:
Replacement:

Round
Rounding base: 500

Weight Noise
Percentage:

- Actual top and bottom value for variable ASSETS.
- Rounding base. **Click Apply.**
- Use of the method is shown in the listbox by an X.

Modify | Modify numerical variables | WeightNoise



Numerical Variables

Modified	Variable
X	WEIGHT
X	INCOME
X	ASSETS
	DEBTS

Bottom Coding
Minimum: 112.7
Bottom Value:
Replacement

Top Coding
Maximum: 420
Top Value:
Replacement:

Round
Rounding base:

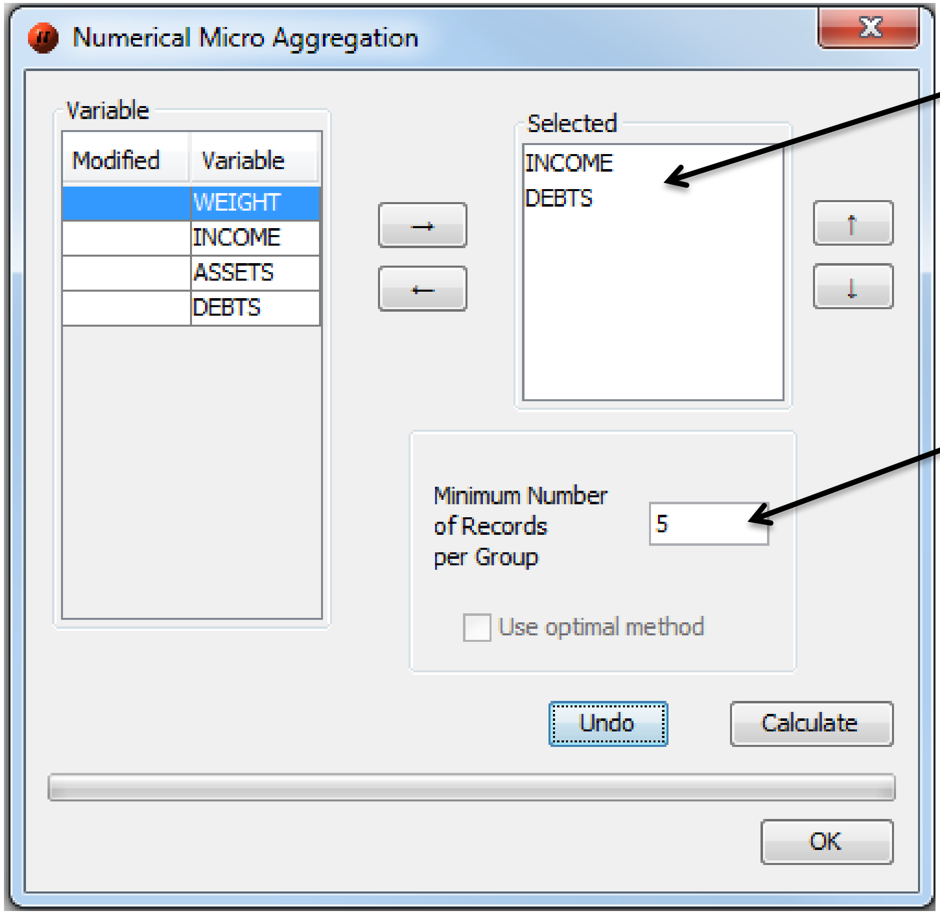
Weight Noise
Percentage: 5

- Actual top and bottom value for variable WEIGHT.

- Percent of the weight noise. **Click Apply.**

- Use of the method is shown in the listbox by an X.

Modify | Numerical Micro Aggregation



Numerical Micro Aggregation

Variable

Modified	Variable
	WEIGHT
	INCOME
	ASSETS
	DEBTS

Selected

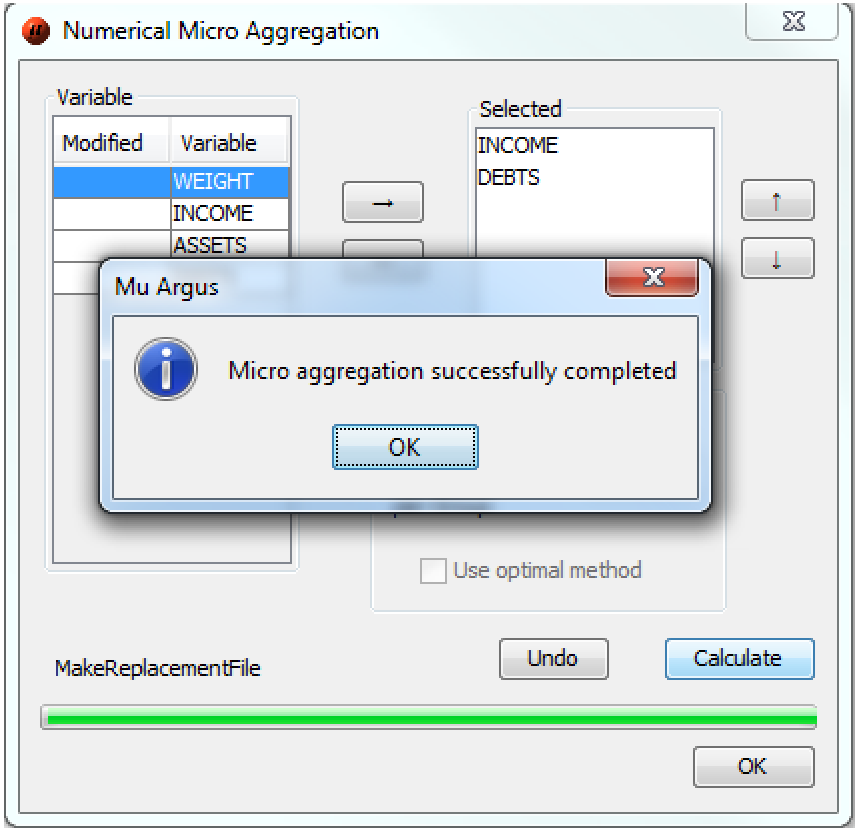
INCOME
DEBTS

Minimum Number of Records per Group: 5

□ Use optimal method

- Selected variables will be „microaggregated“.
- Minimum number of records per group.
- Use of optimal method possible only for a single variable and a small microdata set.

Modify | Numerical MicroAggregation



Selected: INCOME, DEBTS

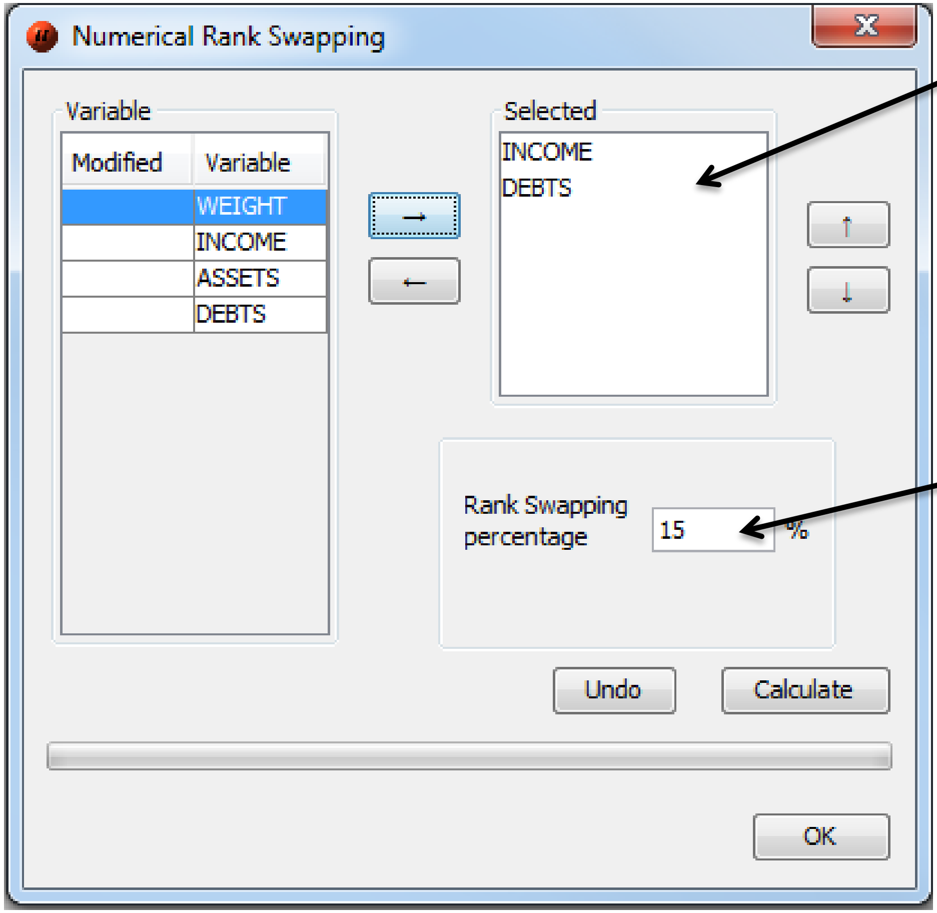
☐ Use optimal method

MakeReplacementFile

Micro aggregation successfully completed

- **Click Calculate → OK → OK.**

Modify | Numerical Rank Swapping



Numerical Rank Swapping

Variable

Modified	Variable
	WEIGHT
	INCOME
	ASSETS
	DEBTS

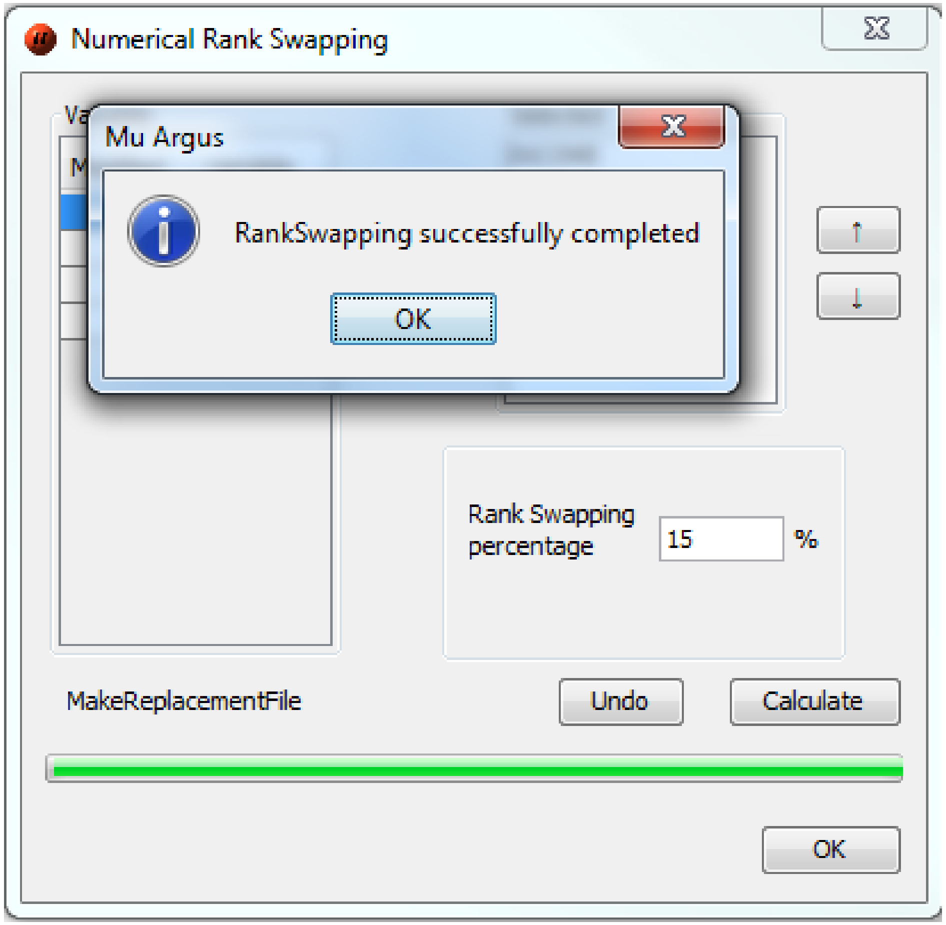
Selected

INCOME
DEBTS

Rank Swapping percentage: 15 %

- The rank swapping will be applied on selected variables.
- Percentage for rank swapping.
- The procedure is applied on each variable individually.

Modify | Numerical Rank Swapping


Numerical Rank Swapping

Mu Argus

RankSwapping successfully completed

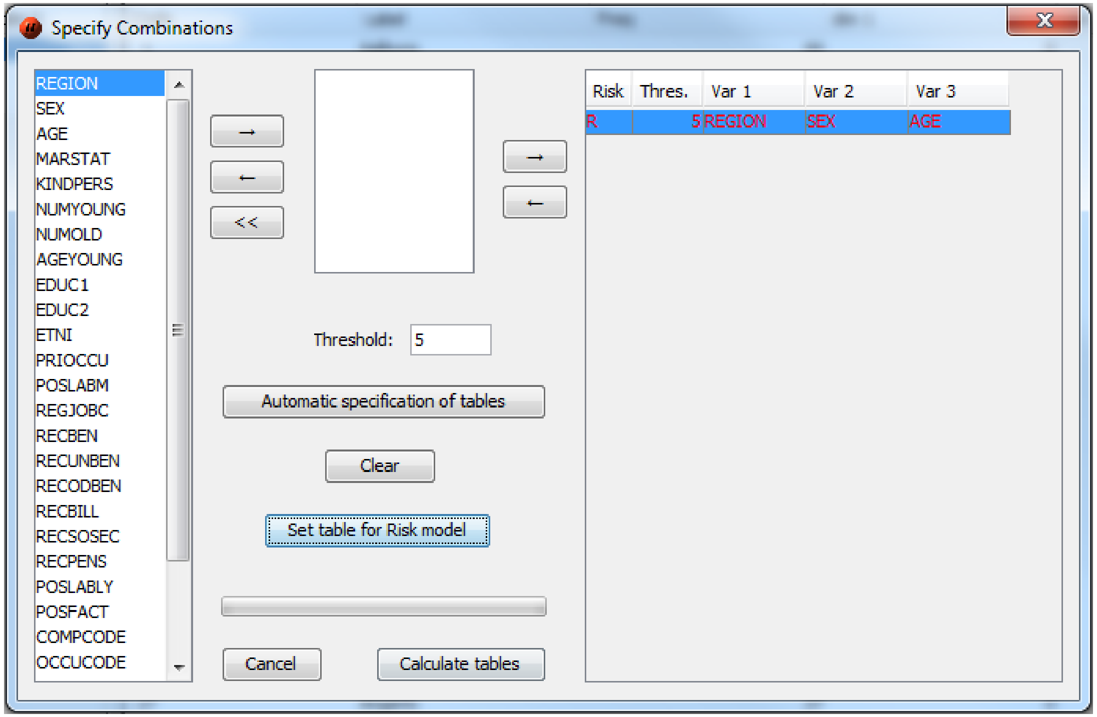
OK

Rank Swapping percentage: 15 %

MakeReplacementFile

- **Click Calculate → OK → OK.**

Specify | Combinations



Threshold: 5

Risk	Thres.	Var 1	Var 2	Var 3
R	5	REGION	SEX	AGE

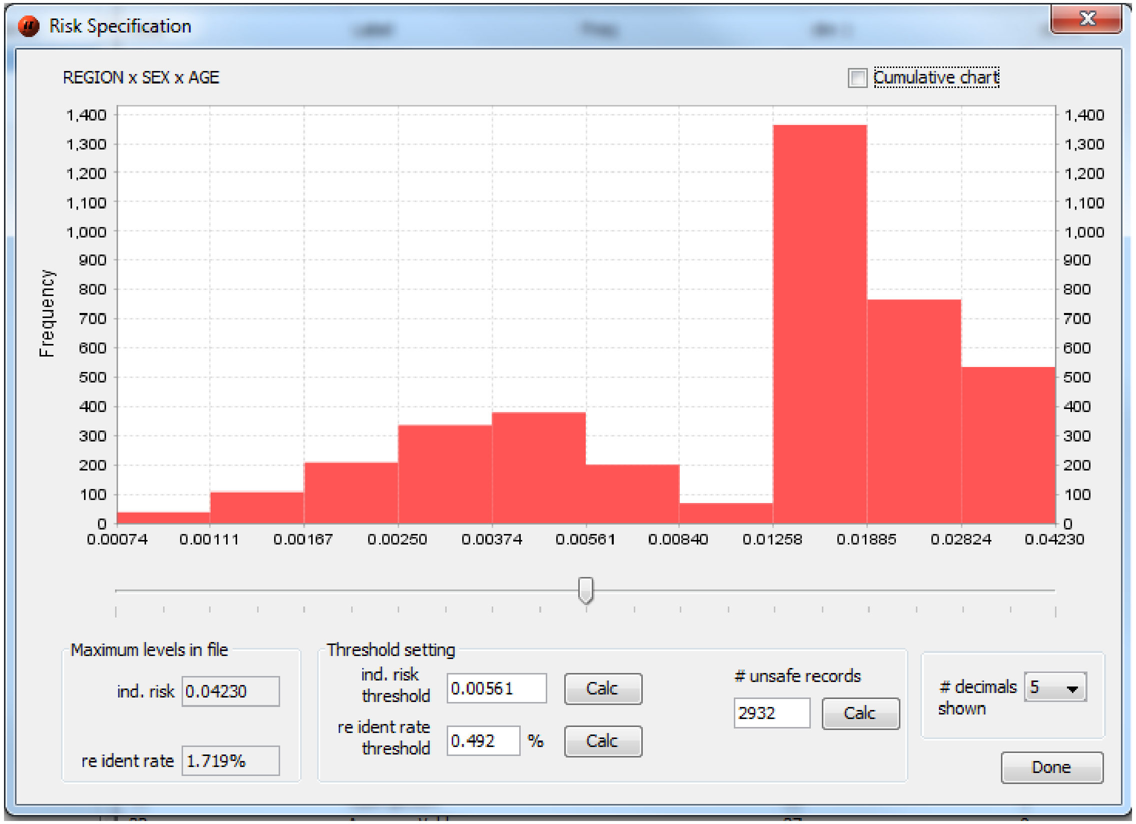
Risk estimation:

- Combination(s) of variables can be selected.
- Click Set table for Risk model (R).
- Click Calculate tables.
- Overlapping risk tables are not allowed.

Modify | … Risk Specification

- No perturbation methods should be used.
- If household id present in microdata → Modify | Household Risk Specification
- If NO household id present in microdata → Modify | Individual Risk Specification

Modify | Individual Risk Specification



Maximum levels in file

ind. risk 0.04230

re ident rate 1.719%

Threshold setting

ind. risk threshold 0.00561

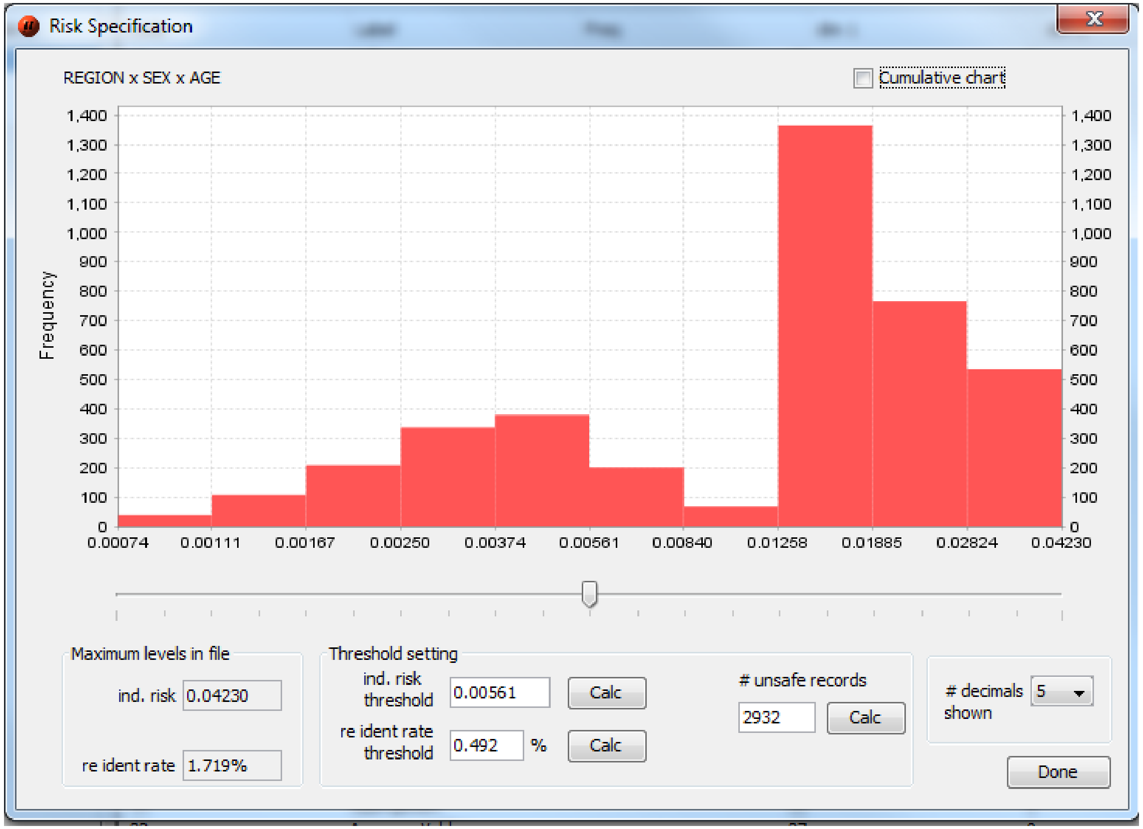
re ident rate threshold 0.492 %

unsafe records 2932

decimals shown 5

- Maximum levels in file:
  - Inv. risk (max)
  - Re ident rate (expected re-identifications)
- Threshold setting:
  - Slider
  - Write a threshold in the „*ind. risk threshold*“ text box.

Modify | Individual Risk Specification



□ Cumulative chart

Maximum levels in file

ind. risk 0.04230

re ident rate 1.719%

Threshold setting

ind. risk threshold 0.00561

re ident rate threshold 0.492 %

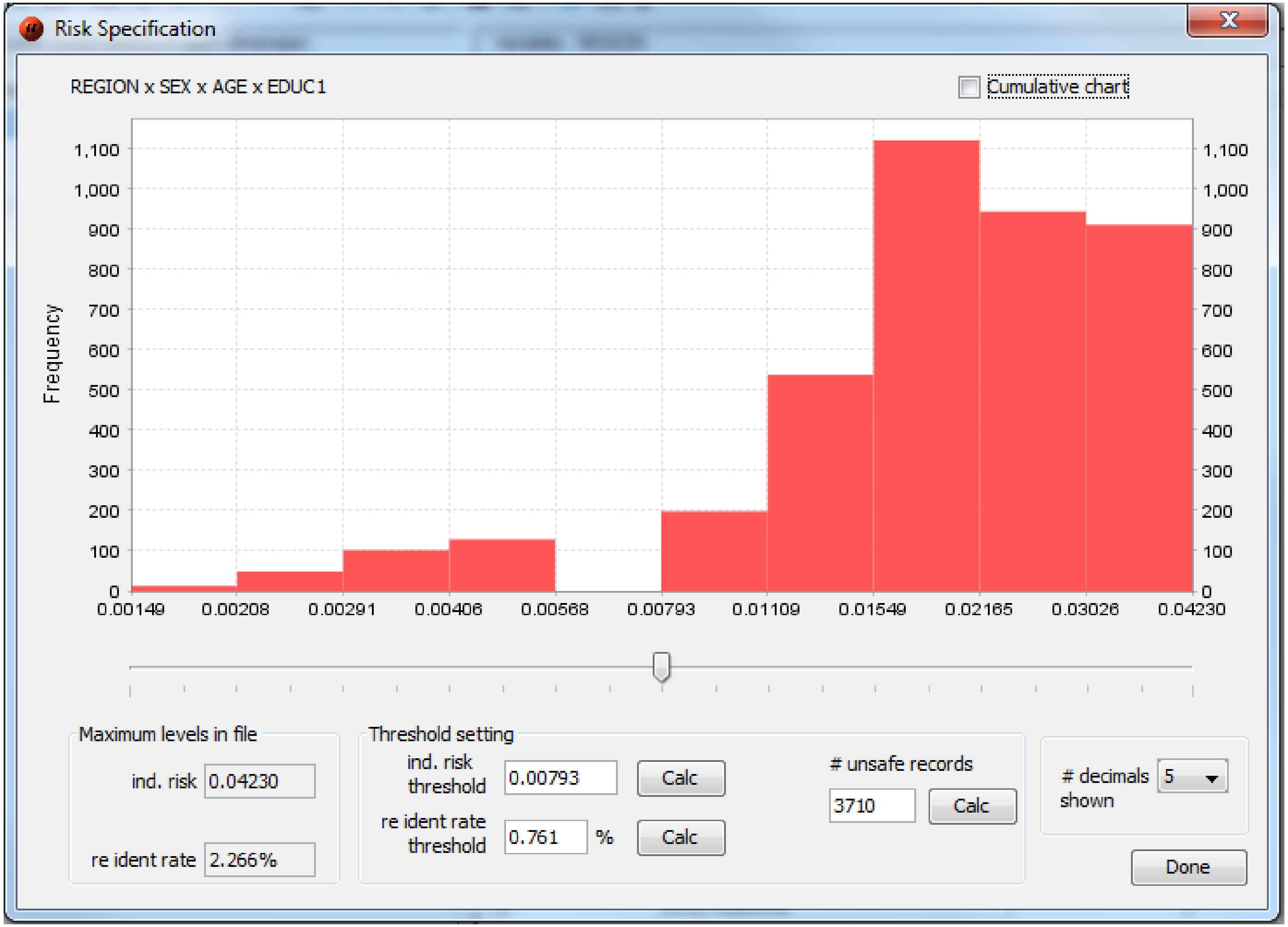
unsafe records 2932

decimals shown 5

- Threshold setting:
  - Inv. Risk threshold ≤ Inv. Risk (max)
  - Re ident rate threshold ≤ Re ident rate
  - Number of unsafe records

- Pressing „Done“ sets the ind. risk threshold → used for local suppression.

Example – before global recode



REGION x SEX x AGE x EDUC1

□ Cumulative chart

Maximum levels in file

ind. risk 0.04230

re ident rate 2.266%

Threshold setting

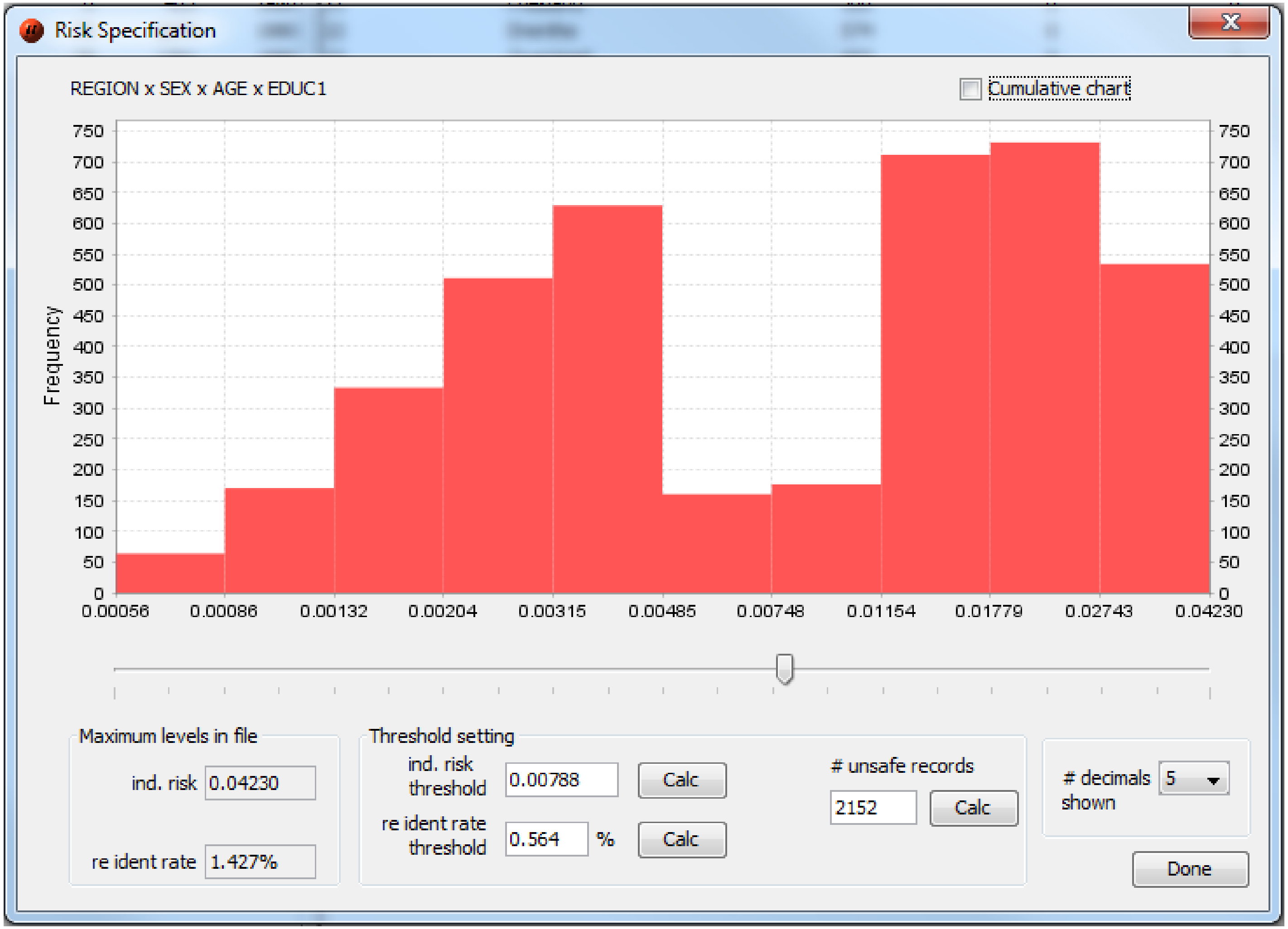
ind. risk threshold 0.00793

re ident rate threshold 0.761 %

unsafe records 3710

decimals shown 5

Example - after global recode



Maximum levels in file

ind. risk 0.04230

re ident rate 1.427%

Threshold setting

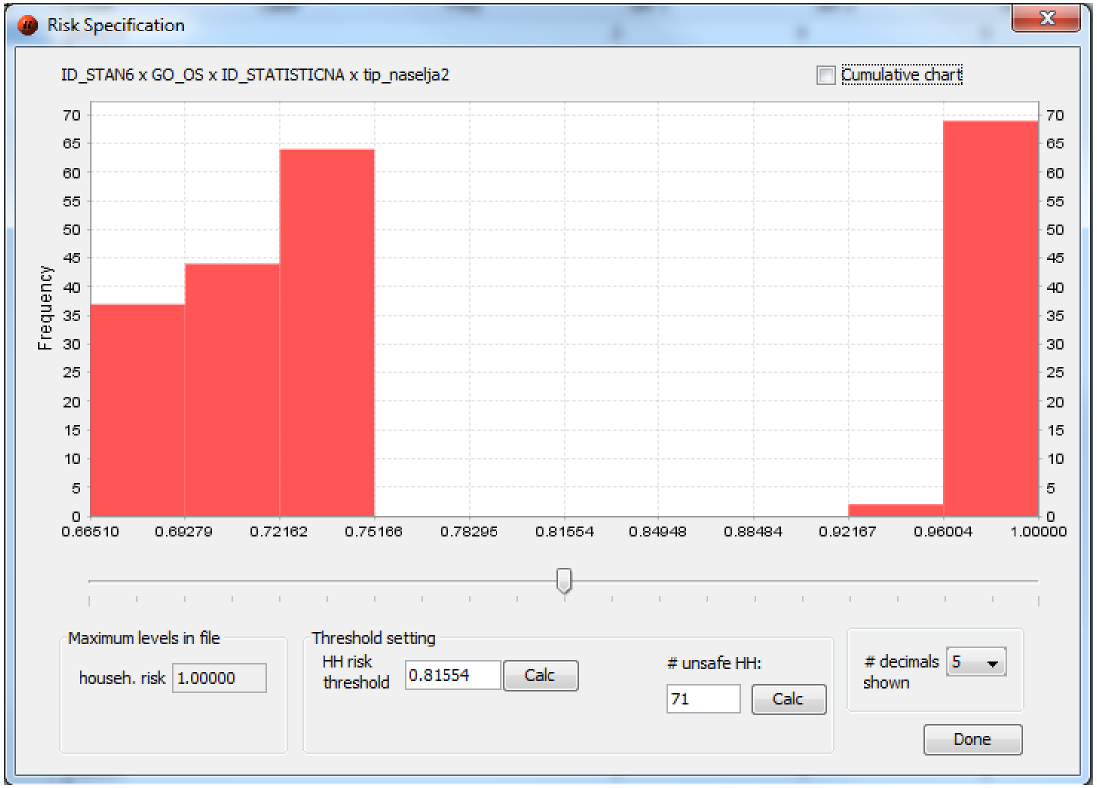
ind. risk threshold 0.00788

re ident rate threshold 0.564 %

unsafe records 2152

decimals shown 5

Modify | Household Risk Specification



Risk Specification

ID_STAN6 x GO_OS x ID_STATISTICNA x tip_naselja2

□ Cumulative chart

Maximum levels in file

househ. risk: 1.00000

Threshold setting

HH risk threshold: 0.81554

unsafe HH: 71

decimals shown: 5

- Threshold setting:
  - Slider
  - Write a threshold in the „*HH risk threshold*“ text box.

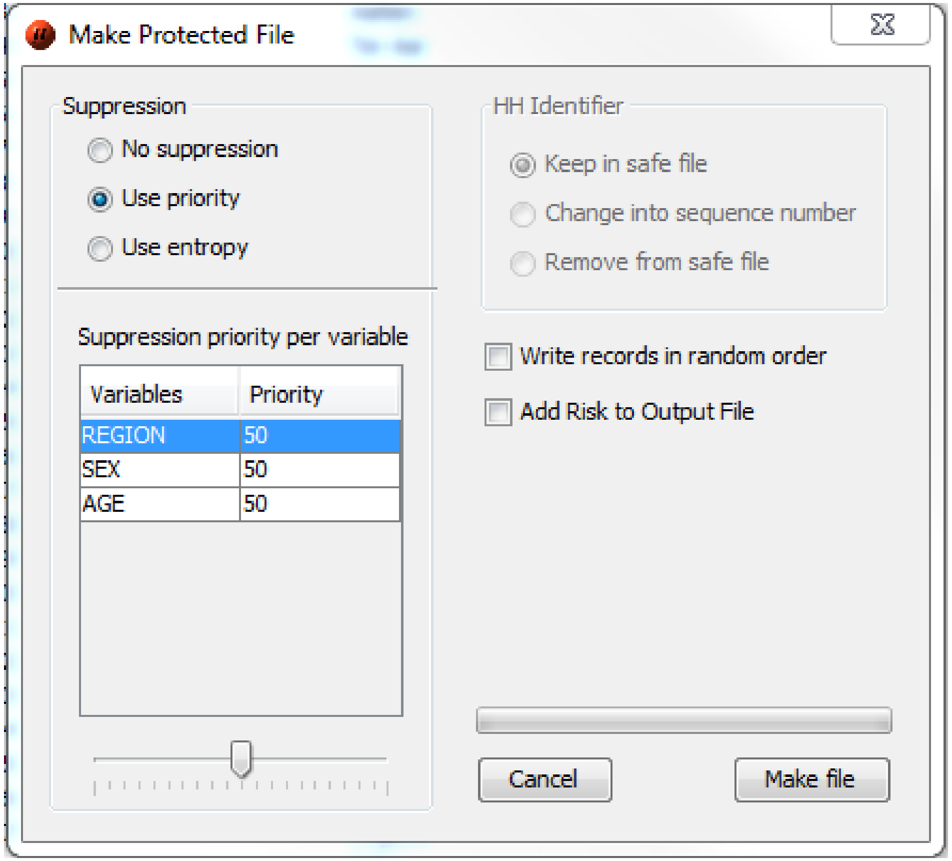
- d is the household size, r_h is the threshold for hh risk. A household member is at risk if the individual risk is higher than or equal to r_h/d.

Modify | … Risk Specification

- The re-identification rate is very often used for determining the threshold for the individual risk
  - E.g. 5 persons out of 4,000 can be identified -> re-identification rate is $\frac{5}{4000} = 0.00125$.

- After determining the records at risk (acceptable information loss) local suppression is used.

Output | Make protected file



Make Protected File

Suppression

○ No suppression

● Use priority

○ Use entropy

Suppression priority per variable

Variables	Priority
REGION	50
SEX	50
AGE	50

HH Identifier

● Keep in safe file

○ Change into sequence number

○ Remove from safe file

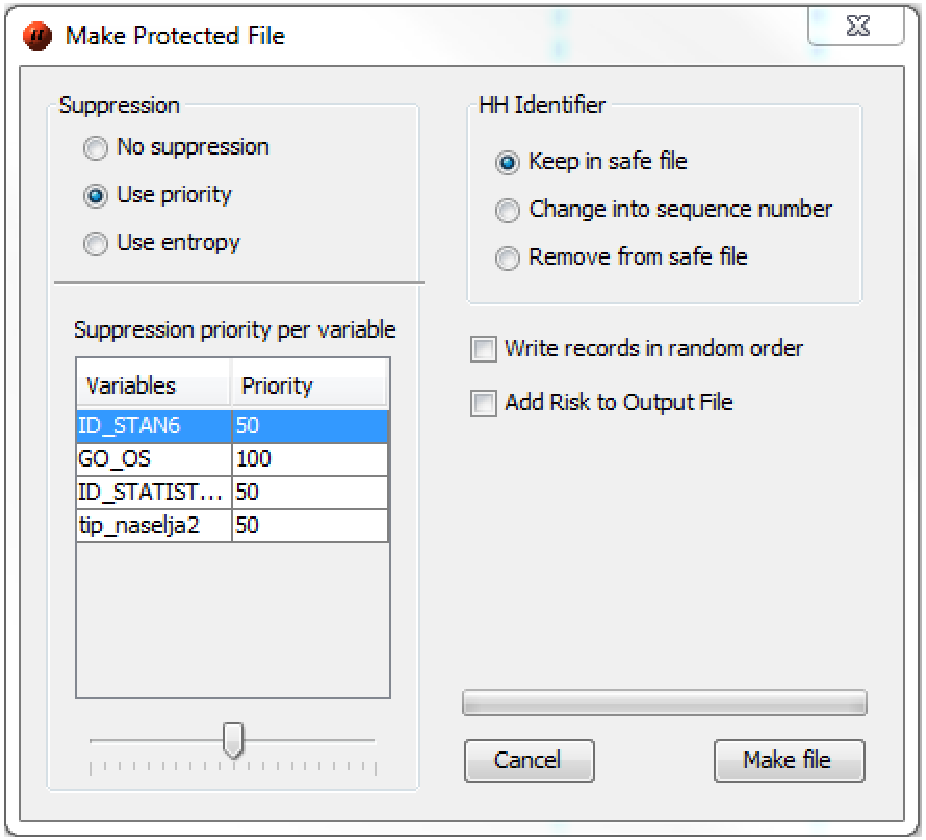
□ Write records in random order

□ Add Risk to Output File

- Local suppression for unsafe combinations
  - **<u>Use priority</u>** (higher value means smaller information loss) **– slider!**
  - **<u>Use entropy</u>** (the variable with the highest number of small categories is suppressed)

- In case of more unsafe combinations for one record, information loss is minimized.

Output | Make protected file



Make Protected File

Suppression
- ○ No suppression
- ● Use priority
- ○ Use entropy

Suppression priority per variable

Variables	Priority
ID_STAN6	50
GO_OS	100
ID_STATIST…	50
tip_naselja2	50

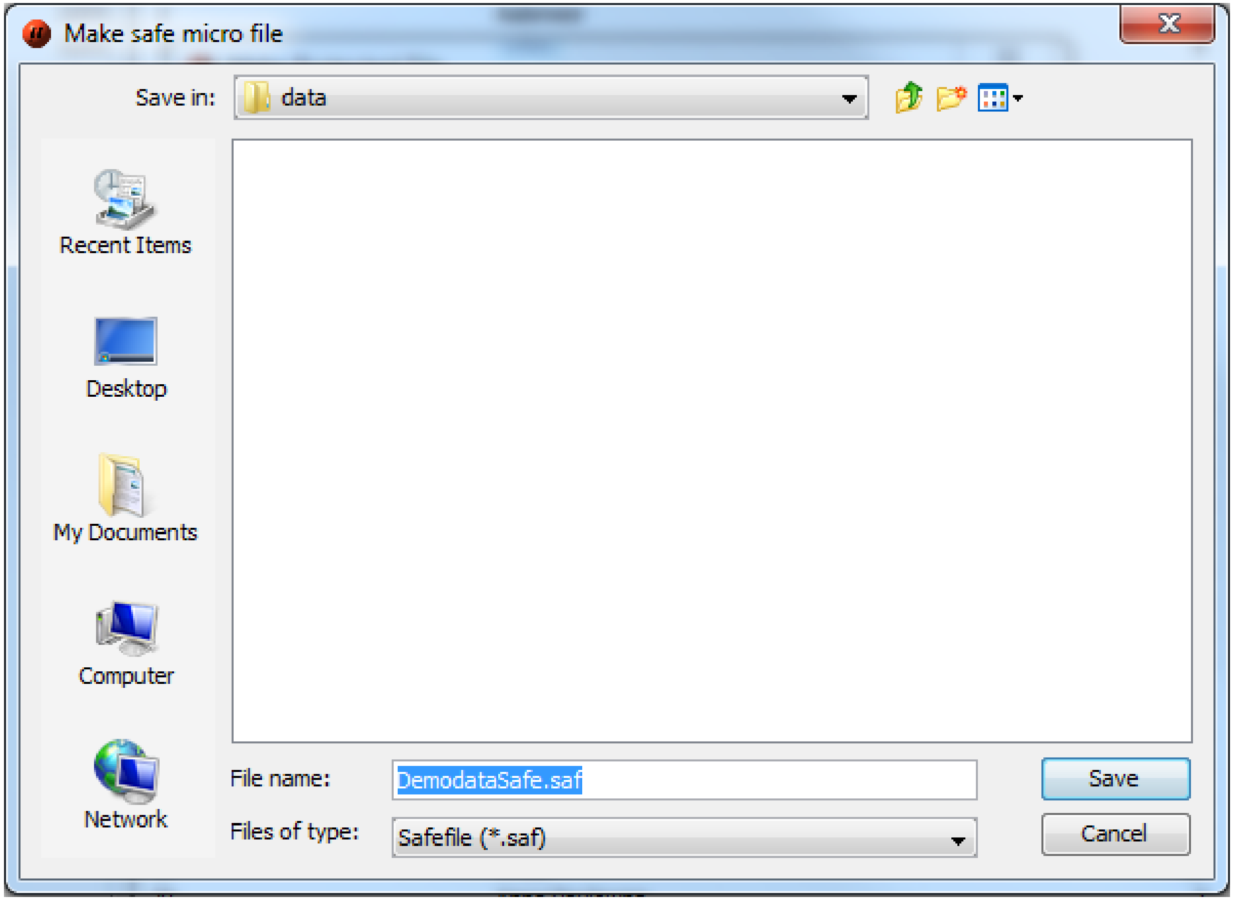
HH Identifier
- ● Keep in safe file
- ○ Change into sequence number
- ○ Remove from safe file

□ Write records in random order

□ Add Risk to Output File

- Household identifier:
  - Do not change.
  - Change it into a simple sequence number.
  - Remove it from the dataset.

Output | Make protected file



Make safe micro file

Save in: data

File name: DemodataSafe.saf

Files of type: Safefile (*.saf)

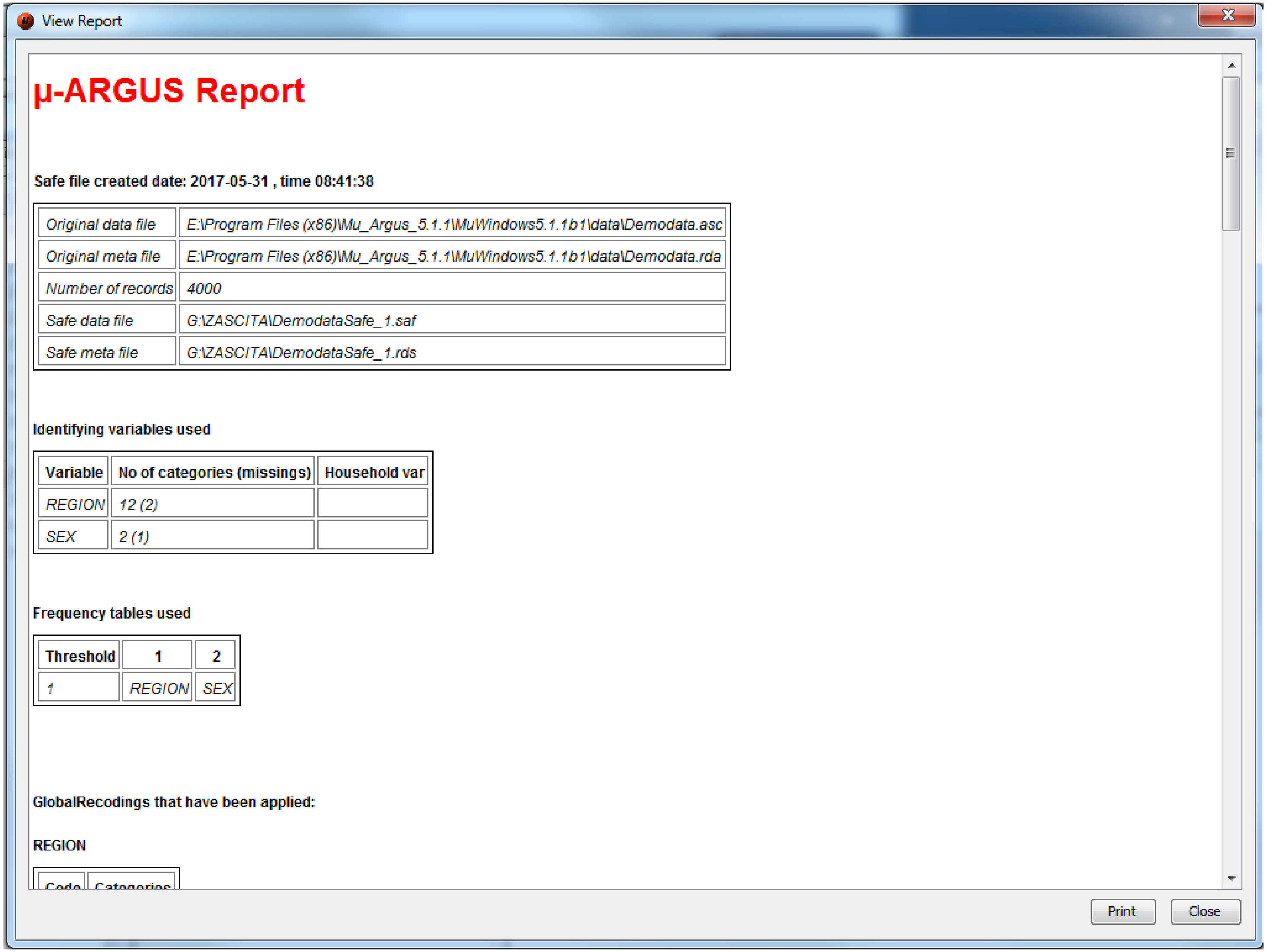
- Click **Make file** -> Choose location for the safe dataset (.saf)

Output | Make protected file

- Two files are created:
  - Metadata file of the safe file (.rds)
  - Safe file (.saf)

- The structure is the same as for input files

Output | View report



μ-ARGUS Report

Safe file created date: 2017-05-31 , time 08:41:38

Original data file	*E:\Program Files (x86)\Mu_Argus_5.1.1\MuWindows5.1.1b1\data\Demodata.asc*
Original meta file	*E:\Program Files (x86)\Mu_Argus_5.1.1\MuWindows5.1.1b1\data\Demodata.rda*
Number of records	*4000*
Safe data file	*G:\ZASCITA\DemodataSafe_1.saf*
Safe meta file	*G:\ZASCITA\DemodataSafe_1.rds*

Identifying variables used

Variable	No of categories (missings)	Household var
REGION	*12 (2)*	
SEX	*2 (1)*	

Frequency tables used

Threshold	1	2
1	*REGION*	*SEX*

GlobalRecodings that have been applied:

REGION

Code	Categories