Disclosure review and de-identification of quantitative data - hands-on with sdcMicro -

Cristina Magder
UK Data Service
University of Essex

Managing and sharing research data: best practice for data protection

London, 28-29 November 2018

UK Data Service

sdcMicro GUI - exercise

The following slides will:

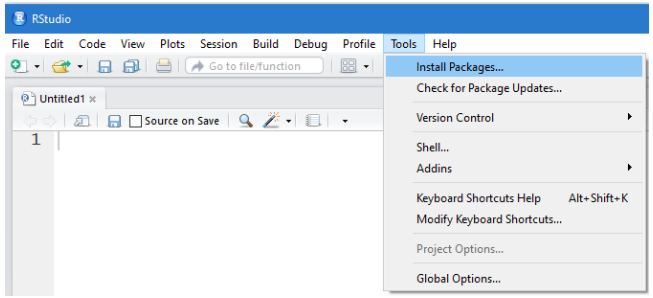
- **provide an overview of the sdcMicro GUI**
- **demonstrate a few features of the GUI**
- **supply the basic knowledge needed to use this SDC Tool**

sdcMicro package

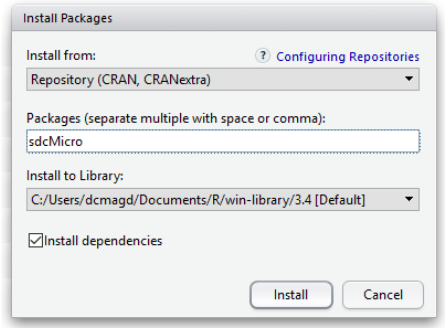
Make sure you have the sdcMicro package installed.

If not:

- in RStudio click on the **Tools Tab** in the upper toolbar and select **Install Packages**



- then **type in sdcMicro** (pay attention to spelling and lower and upper letters) and **click on Install**




```
package 'sdcMicro' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
```

UK Data Service

How to open sdcMicro GUI

```
> library(sdcMicro)
--------
This is sdcMicro v5.2.0.
For references, please have a look at citation('sdcMicro')
Note: since version 5.0.0, the graphical user-interface is a shiny-app that can be started
 with sdcApp().
Please submit suggestions and bugs at: https://github.com/sdcTools/sdcMicro/issues
--------
Warning message:
package 'sdcMicro' was built under R version 3.4.4
> sdcApp()
Warning: package 'shiny' was built under R version 3.4.3
Warning: package 'rhandsontable' was built under R version 3.4.3
Warning: package 'haven' was built under R version 3.4.3

Attaching package: 'DT'

The following objects are masked from 'package:shiny':

    dataTableOutput, renderDataTable

Warning: package 'data.table' was built under R version 3.4.3
data.table 1.10.4.3
  The fastest way to learn (by data.table authors): https://www.datacamp.com/courses/data-
analysis-the-data-table-way
  Documentation: ?data.table, example(data.table) and browseVignettes("data.table")
  Release notes, videos and slides: http://r-datatable.com
```

UK Data Service

sdcApp

This graphical user interface of `sdcMicro` allows you to anonymize microdata even if you are not an expert in the `R` programming language. Detailed information on how to use this graphical user-interface (GUI) can be found in a tutorial (a so-called vignette) that is included in the `sdcMicro` package. The vignette is available on GitHub pages and via the CRAN website. The vignette can also be viewed offline by typing `vignette("sdcMicro", package="sdcMicro")` into your `R` prompt.

For information on the support and development of the graphical user interface, please click here .

Getting started

To get started, you need to upload a file with microdata to the GUI. You can do so by clicking this button. Alternatively, you can upload a previously saved problem instance by clicking here.

Set storage path

Currently, all output, such as anonymized data, scripts and reports, will be saved to `C:/Users/dcmagd/Documents` .

You can change the default path, where all output from the GUI will be saved. You can change this path any time later as well by returing to this tab.

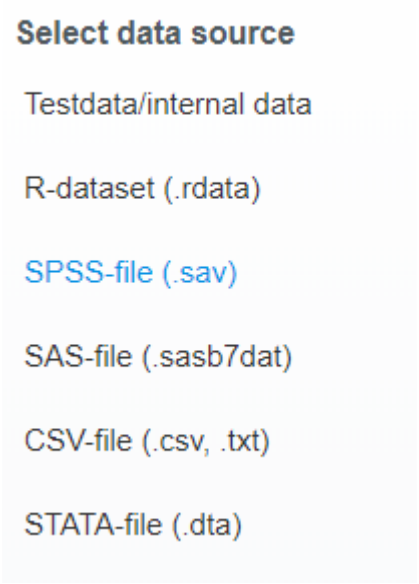
Enter a directory where any exported files (data, script, problem instances) should be saved to

e.g: C:/Users/dcmagd/Documents

sdcMicro GUI has 7 main menus:

- About/Help
- Microdata
- Anonymize
- Risk/Utility
- Export Data
- Reproducibility
- Undo

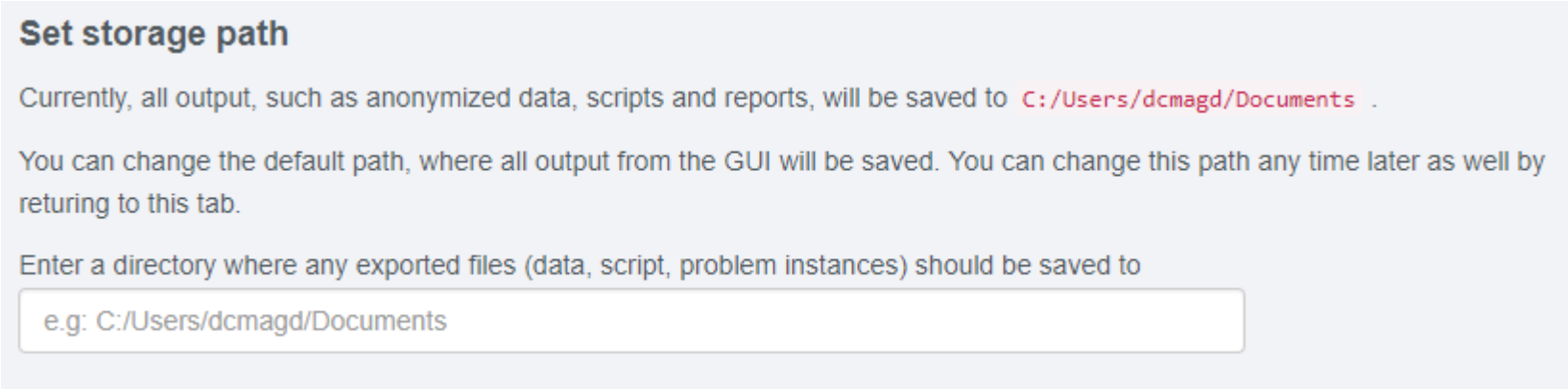
All menus have various options displayed on the left-hand side such as:



UK Data Service

In the first Menu **About/Help** you can:

- Set the path for storage (also the default one is provided)



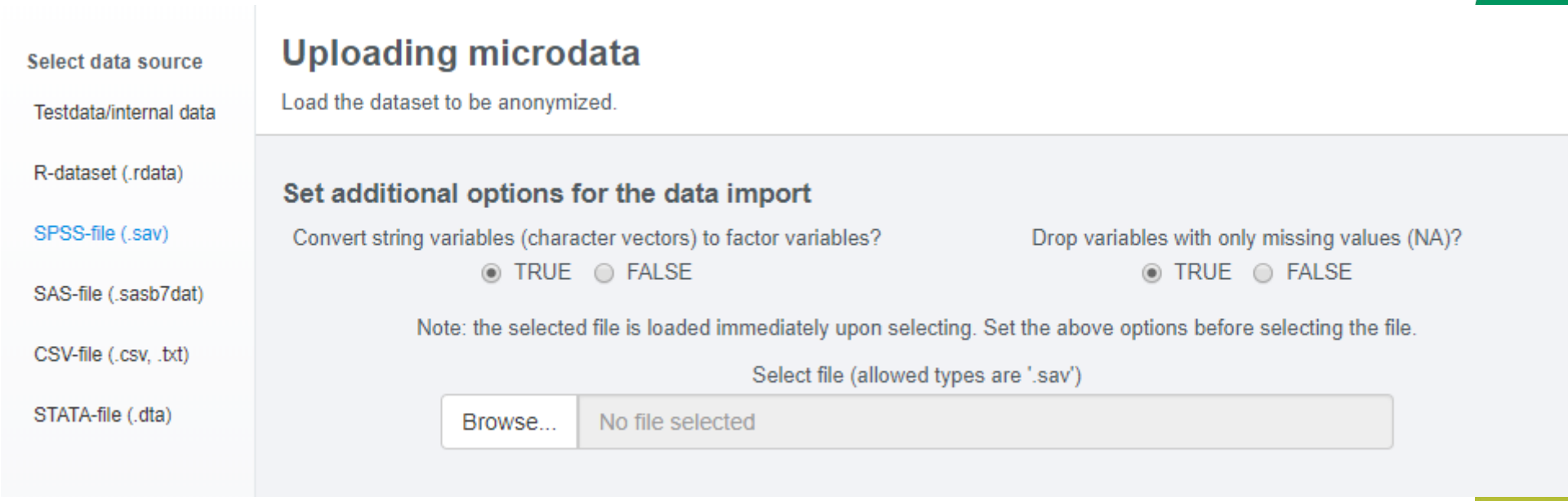
- Stop the interface
- Restart the interface
- Contact and Feedback (takes you directly to the GitHub dedicated page where issues and bugs are reported)

UK Data Service

Importing Data

About/Help | Microdata | Anonymize | Risk/Utility | Export Data | Reproducibility | Undo

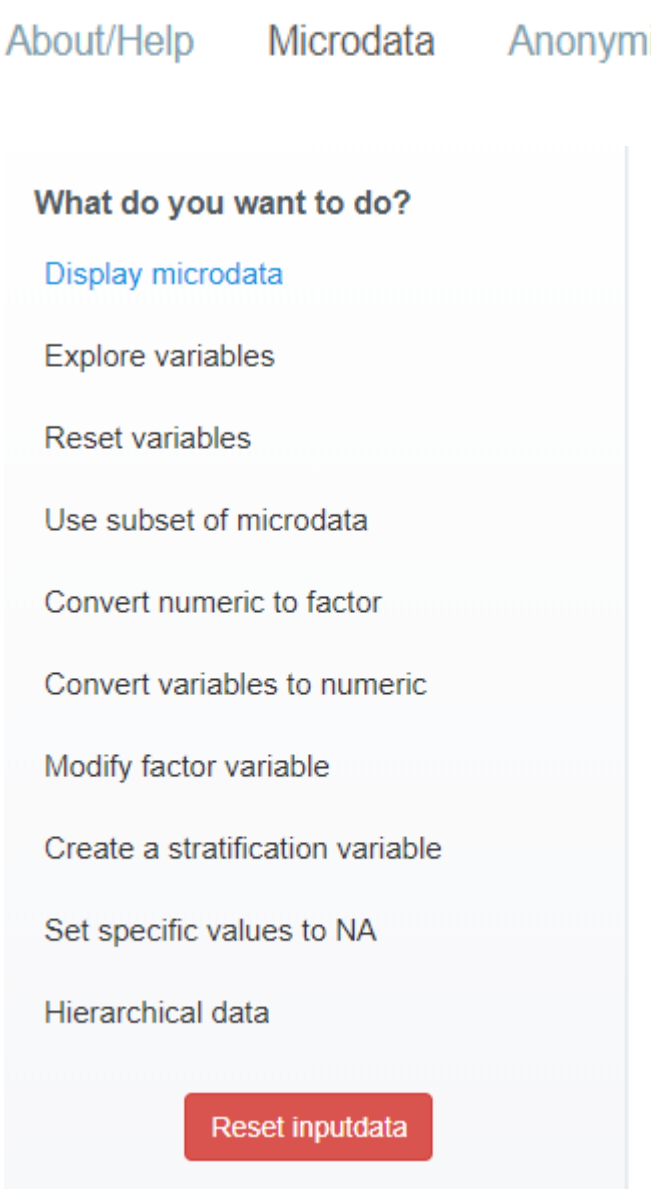
Can import 5 types of data files



Depending on the format your data is in select the corresponding tab on the left hand side (eg for test data SPSS) and click on Browse to find and load your data.

UK Data Service

Importing Data – once file has been loaded

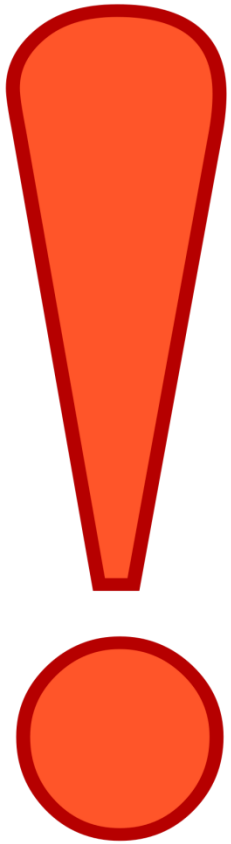


Make sure your data is in the correct format once imported by exploring the variables.

If needed use **Convert numeric to factor** or **Convert variables to numeric** (eg variable Age might have a label for value 99 and R will think Age is a factor when is actually a numeric variable). Not defining variables correctly will have a high impact on calculating risk

If your **key variables have any NA such as Don't know/Refused/Can't remember make sure to define those values as NA in the GUI interface.**

UK Data Service



Always double check

- **Variables formats**
- **NAs**

Once you are happy with your variables you can proceeded to the next step

UK Data Service

Time to set up our SDC problem

Choose your key variables carefully (selecting either categorical or continuous depending on the case and ticking the Weight box for the weight variable)

Anonymize

Select variables and set parameters to create the SDC problem.

Select variables

Variable name	Type	Key variables	Weight	Hierarchical identifier	PRAM	Delete	Number of levels	Number of missing
rsex	factor	No / Cat. / Cont.					2	0
rage	numeric	No / Cat. / Cont.					68	0
rethnic	factor	No / Cat. / Cont.					13	0
relig	factor	No / Cat. / Cont.					6	0
highqual	factor	No / Cat. / Cont.					7	0
occup	factor	No / Cat. / Cont.					23	0
cancer	factor	No / Cat. / Cont.					2	0
car	factor	No / Cat. / Cont.					2	0
weight	numeric	No / Cat. / Cont.	✓				88	0
gor	factor	No / Cat. / Cont.					12	0

Setup SDC problem

View/Analyze existing sdcProblem

Show summary

Explore variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PRAM (simple)

PRAM (expert)

Supress values with high risks

Anonymize numerical variables

Top/bottom coding

Microaggregation

Adding noise

Rank swapping

Reset SDC problem

Summary of dataset and variable selection

The loaded dataset consists of `3714` records and `12` variables.

Categorical key variable(s): `rsex` `rethnic` `relig` `highqual` `occup` `gor`

Numerical key variable(s): `rage`

Sampling weight: `weight`

Computation time

The current computation time was ~ `0.5 seconds` .

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level
rsex	2 (2)	1857.000 (1857.000)	1672 (1672)
rethnic	13 (13)	218.471 (218.471)	0 (0)
relig	6 (6)	619.000 (619.000)	12 (12)
highqual	7 (7)	464.250 (464.250)	0 (0)
occup	23 (23)	148.560 (148.560)	0 (0)
gor	12 (12)	309.500 (309.500)	181 (181)

Risk measures for categorical key variables

We expect `44.14` (`1.19%`) re-identifications in the population, as compared to `44.14` (`1.19%`) re-identifications in the original data.

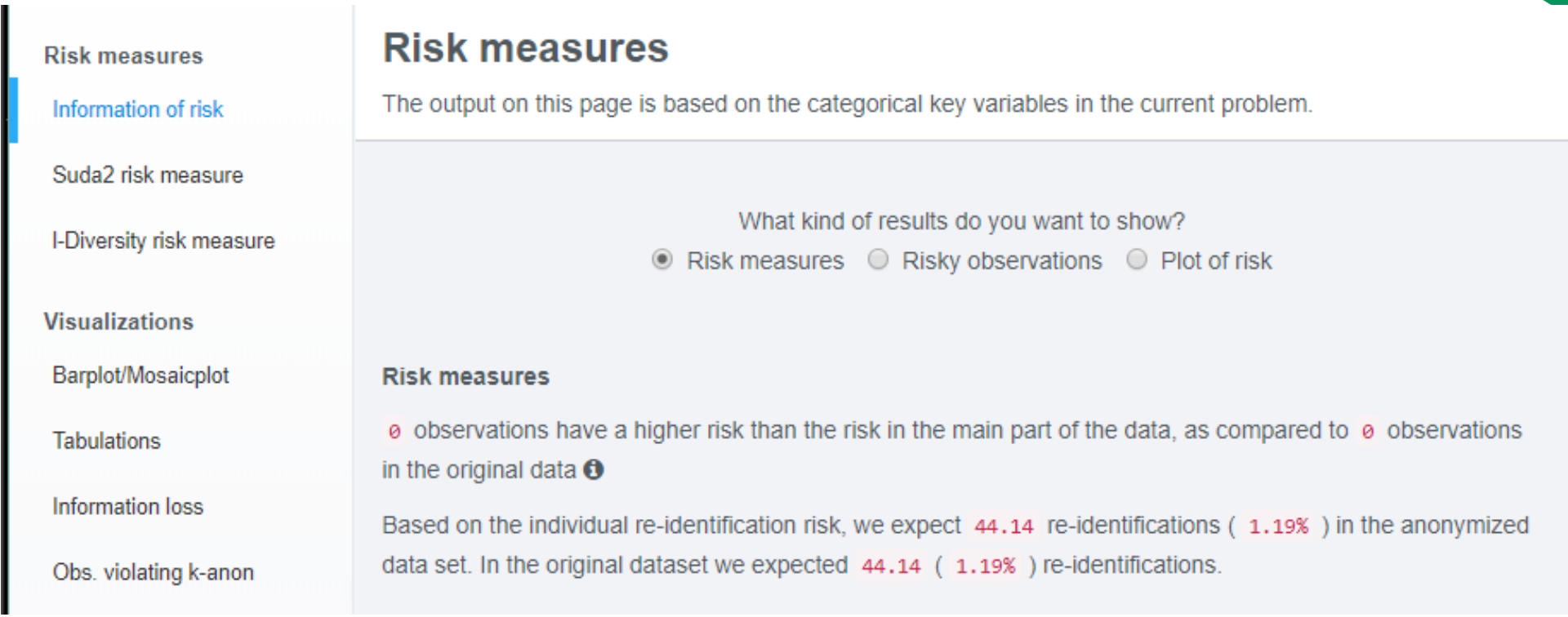
`0` observations have a higher risk than the risk in the main part of the data, as compared to `0` observations in the original data.

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

Once you are familiar with the summary it's time to get an idea about the risk

The **Risk/Utility** provides various sub-menus on the left-hand side, but the first think to inspect is the **Information of Risk**



Risk measures

Information of risk

Suda2 risk measure

I-Diversity risk measure

Visualizations

Barplot/Mosaicplot

Tabulations

Information loss

Obs. violating k-anon

Numerical risk measures

Compare summary statistics

Disclosure risk

Information loss

Risk measures

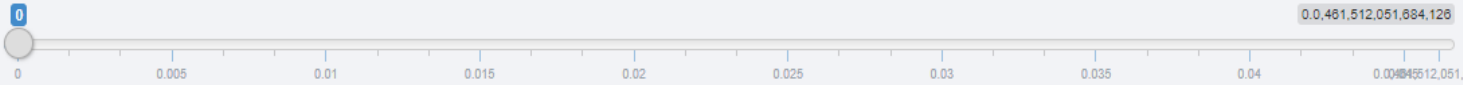
The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

○ Risk measures ◉ Risky observations ○ Plot of risk

Display risky observations in a table

Minimum risk for to be shown in the table



3714 (100.00%) records have a risk larger than 0 .

Show 10 entries

	rsex	rethnic	relig	highqual	occup	gor	fk	Fk	indivRisk
1	Male	Black / African / Caribbean / Black British - African	Christianity	GCSE (D-E)	Accountant	NI	1	135	0.036607
2	Male	Black / African / Caribbean / Black British - African	Islam	Undergraduate	Ambulance Officer	NI	2	234	0.008267
3	Male	Black / African / Caribbean / Black British - African	Hinduism	GCSE (D-E)	Baker	London	1	112	0.042509
4	Male	Black / African / Caribbean / Black British - African	Buddhism	A-levels	Building Inspector	London	1	143	0.034950
5	Female	Black / African / Caribbean / Black British - African	No religion	Postgraduate - PHD	Cardiologist	London	2	334	0.005838

Showing 1 to 10 of 3,714 entries

Previous 1 2 3 4 5 … 372 Next

Risk measures
- Information of risk
- Suda2 risk measure
- I-Diversity risk measure

Visualizations
- Barplot/Mosaicplot
- Tabulations
- Information loss
- Obs. violating k-anon

Numerical risk measures
- Compare summary statistics
- Disclosure risk
- Information loss

Risk measures

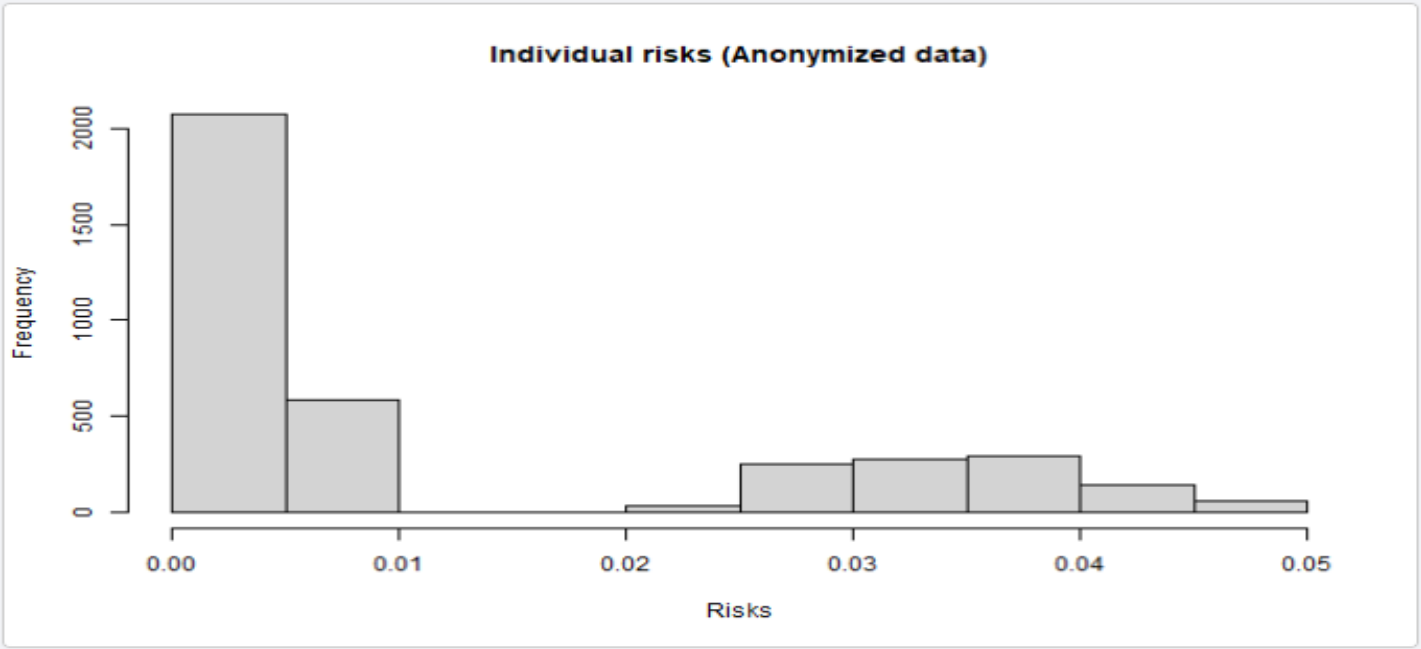
The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

○ Risk measures ○ Risky observations ◉ Plot of risk

Plot showing distribution of individual re-identification risk levels

Anonymized data



Time to see which how each key variables is affecting the risk

Risk measures
- Information of risk
- Suda2 risk measure
- I-Diversity risk measure

Visualizations
- Barplot/Mosaicplot
- Tabulations
- Information loss
- Obs. violating k-anon

Numerical risk measures
- Compare summary statistics
- Disclosure risk
- Information loss

SUDA2 risk measure

The SUDA algorithm is used to search for Minimum Sample Uniques (MSU) in the data among the sample uniques to determine which sample uniques are also specia
more information on SUDA scores.

Reset to choose a different sampling fraction parameter

Suda scores (sampling fraction is 0.1)

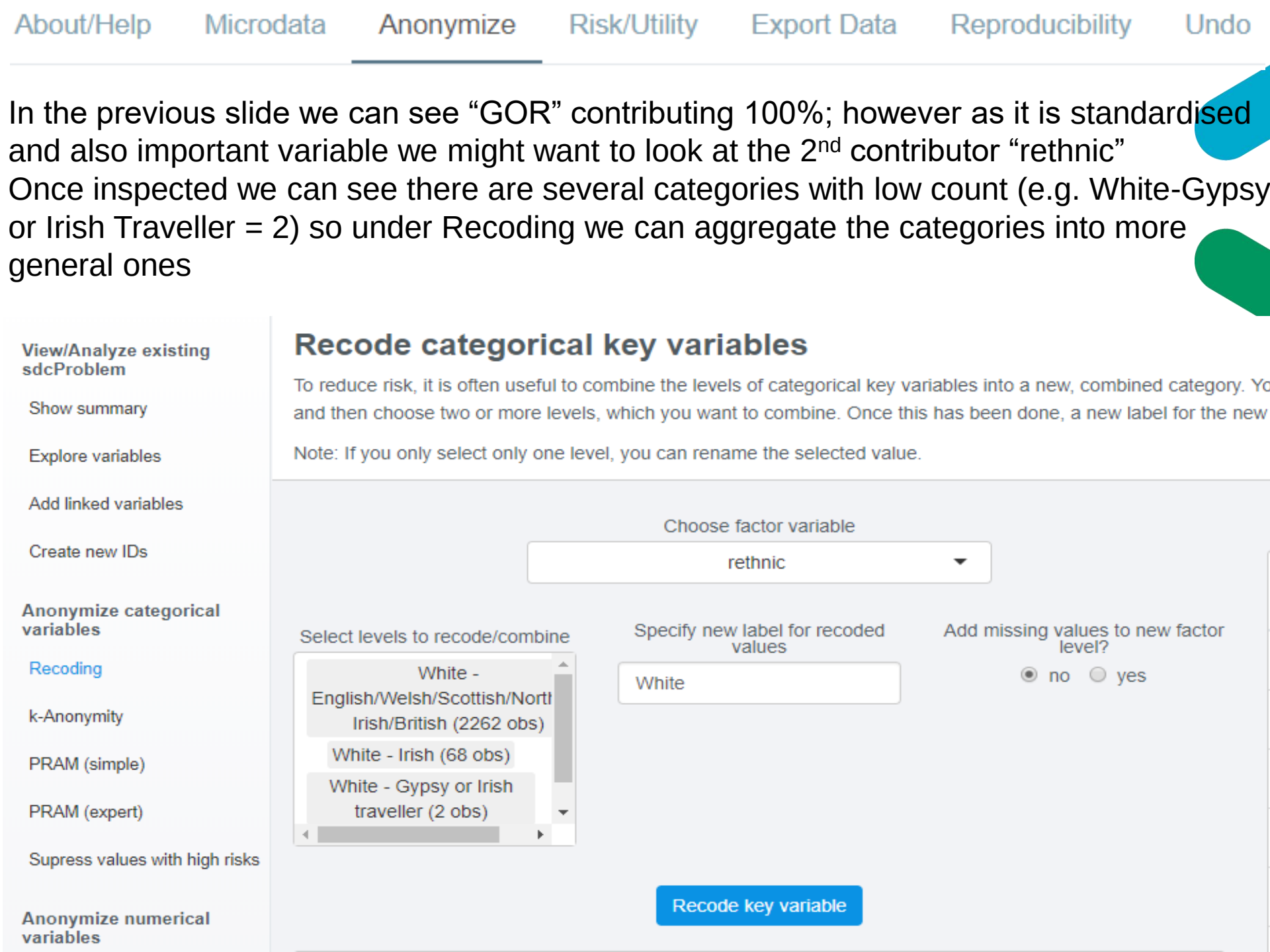
The table below shows the frequencies of the records with a suda score in the specified intervals.

Interval	Number of records
== 0	2655
(0.0, 0.1]	335
(0.1, 0.2]	106
(0.2, 0.3]	207
(0.3, 0.4]	140
(0.4, 0.5]	139
(0.5, 0.6]	54
(0.6, 0.7]	59
> 0.7	19

Attribute contributions

The table below shows the contribution of each categorical key variable to the SUDA scores. The contribution of a variable is the percentage of the total MSUs in the file that include this variable.

variable	contribution
rsex	21.17
rethnic	67.97
relig	32.90
highqual	23.83
occup	60.02
gor	100.00


About/Help
Microdata
Anonymize
Risk/Utility
Export Data
Reproducibility
Undo
In the previous slide we can see “GOR” contributing 100%; however as it is standardised and also important variable we might want to look at the 2nd contributor “rethnic”
Once inspected we can see there are several categories with low count (e.g. White-Gypsy or Irish Traveller = 2) so under Recoding we can aggregate the categories into more general ones
View/Analyze existing sdcProblem
Show summary
Explore variables
Add linked variables
Create new IDs
Anonymize categorical variables
Recoding
k-Anonymity
PRAM (simple)
PRAM (expert)
Supress values with high risks
Anonymize numerical variables
Recode categorical key variables
To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. Yo
and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new
Note: If you only select only one level, you can rename the selected value.
Choose factor variable
rethnic
Select levels to recode/combine
White - English/Welsh/Scottish/North Irish/British (2262 obs)
White - Irish (68 obs)
White - Gypsy or Irish traveller (2 obs)
Specify new label for recoded values
White
Add missing values to new factor level?
no
yes
Recode key variable

Variables from the SDC Perspective

Original data
17 categories

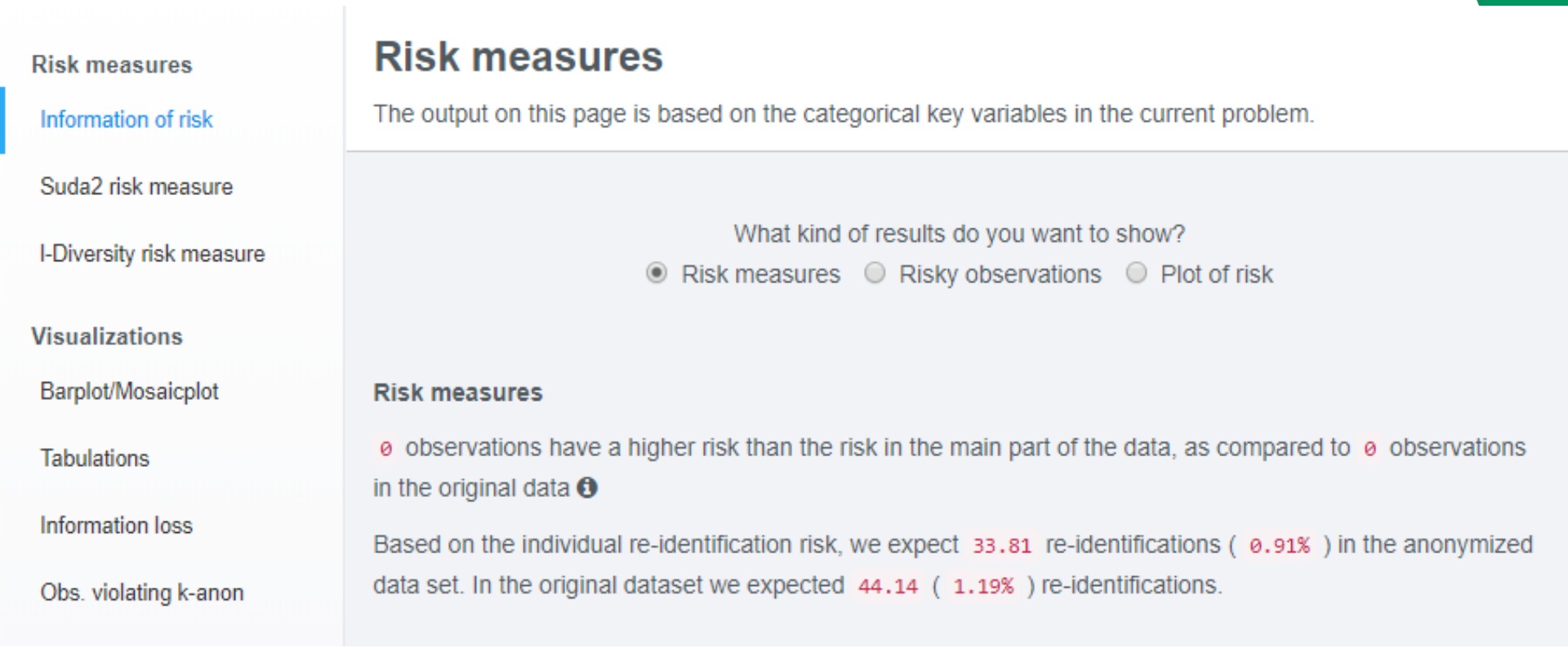
Modified data
5 categories

Original data
Lowest count 1

Modified data
Lowest count 62

UK Data Service

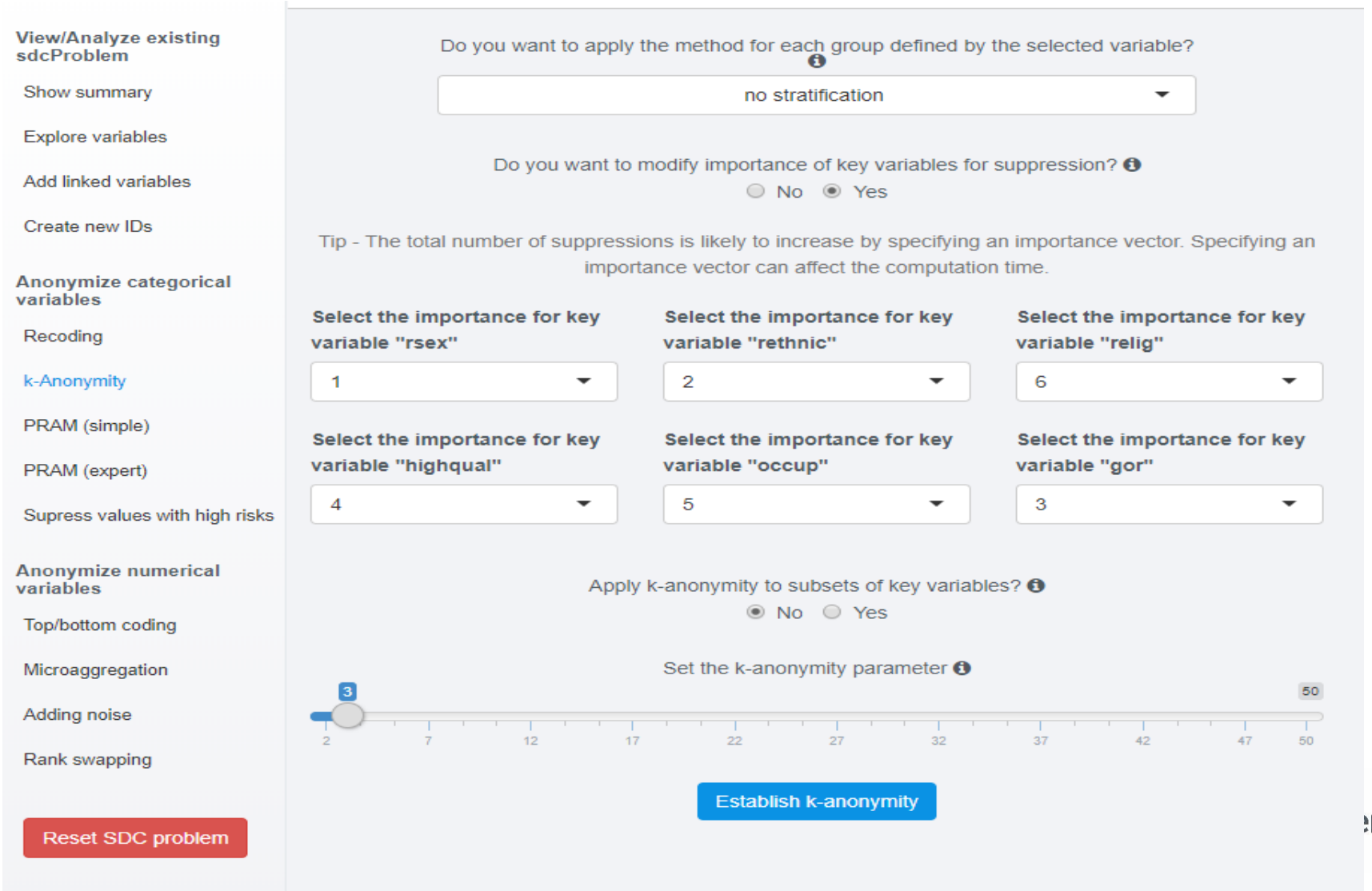
From **1.19% risk to 0.91% risk**
This might seem negligible but it is actuality a **23.5% decrease** (highly significant)



UK Data Service

Let's try to k-anonymise our data

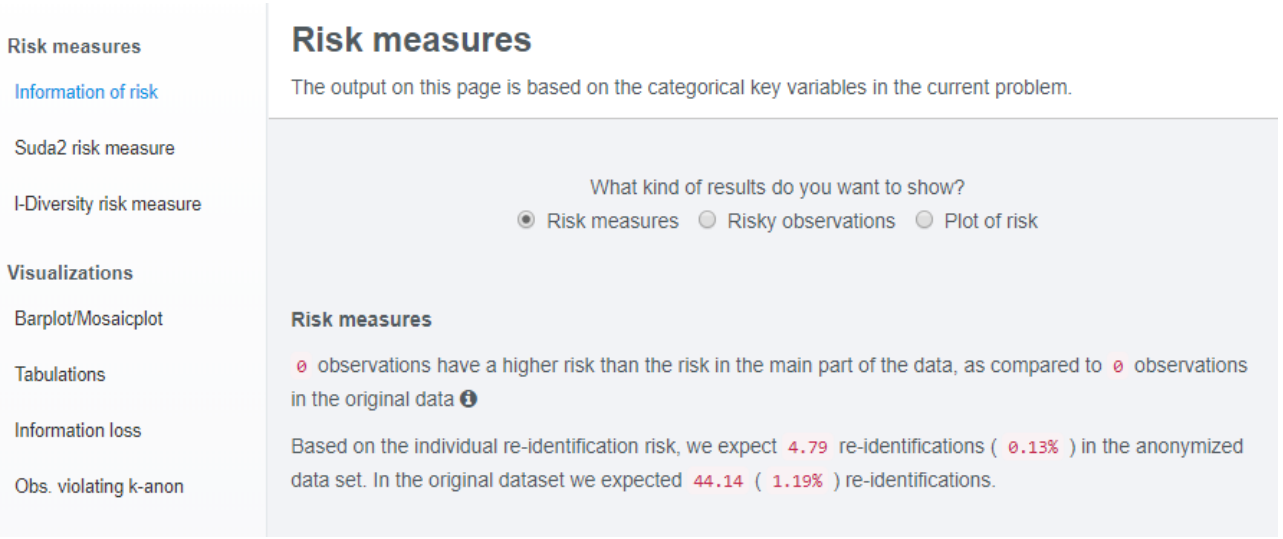
By selecting **yes** on "Do you want to modify the importance for suppression" we gain more control over the values that will be supressed more (**smaller the number higher the importance**, so eg rsex will be the least supressed one)



Information on local suppression

Key variable	Additional suppressions due to last run of kAnon()	Total number of missing values (NA) in variable
rsex	2 (0.054%)	2 (0.054%)
rethnic	10 (0.269%)	10 (0.269%)
relig	643 (17.313%)	643 (17.313%)
highqual	51 (1.373%)	51 (1.373%)
occup	545 (14.674%)	545 (14.674%)
gor	127 (3.419%)	127 (3.419%)

Under Anonymize-> Show Summary we can see how many values for each variables have been supressed



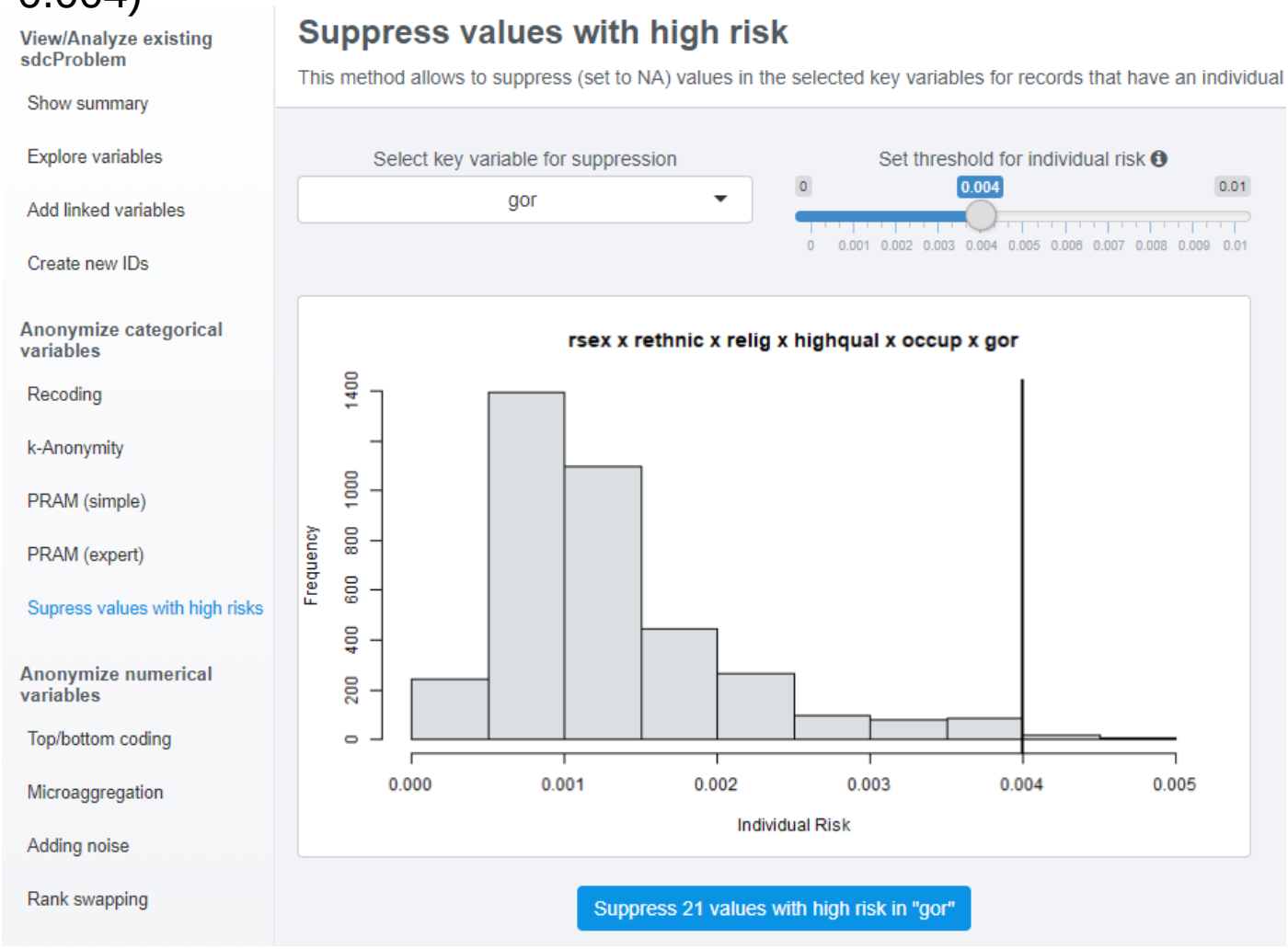
1st Step: Recoding 1.19% to 0.91%

2nd Step: K-Anonymization 0.91% to 0.13

UK Data Service

Furthermore, we could supress values with high risk just for one variable

In our example we are supressing all values with higher risk in GOR (over 0.004)



UK Data Service

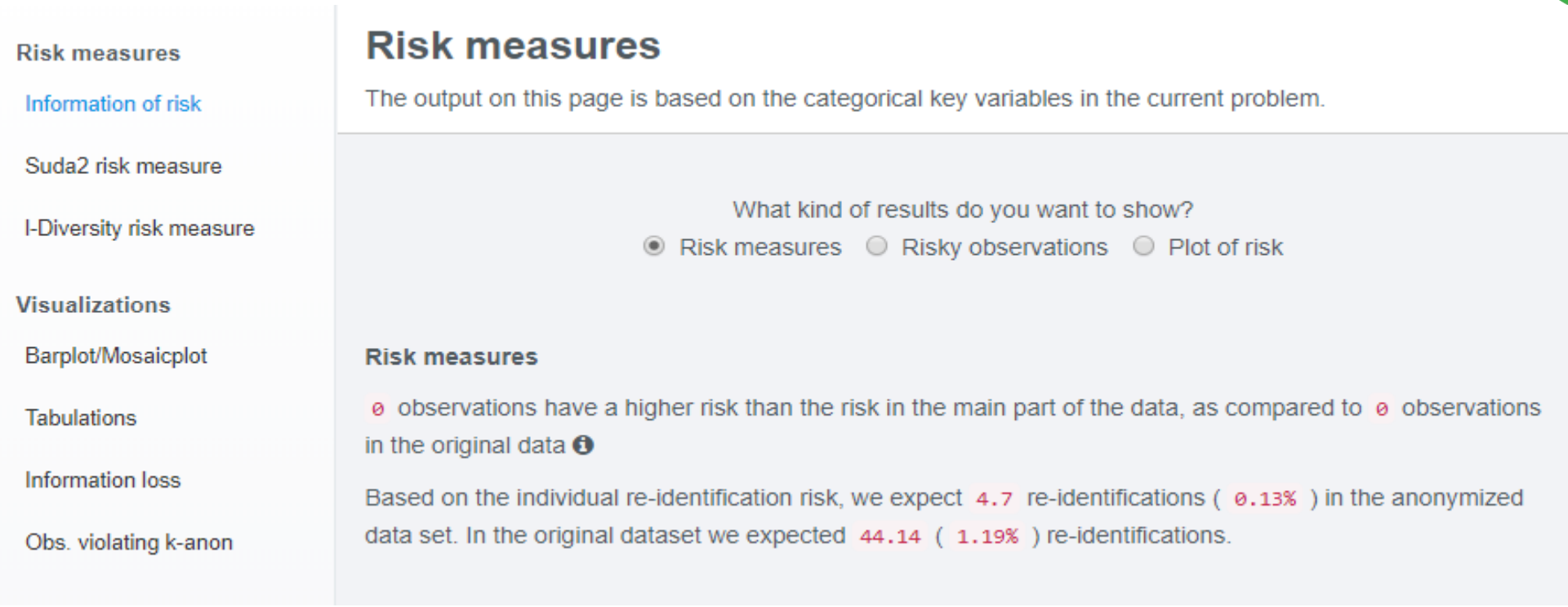
Was it worth it?

1st Step: Recoding
1.19% to 0.91%

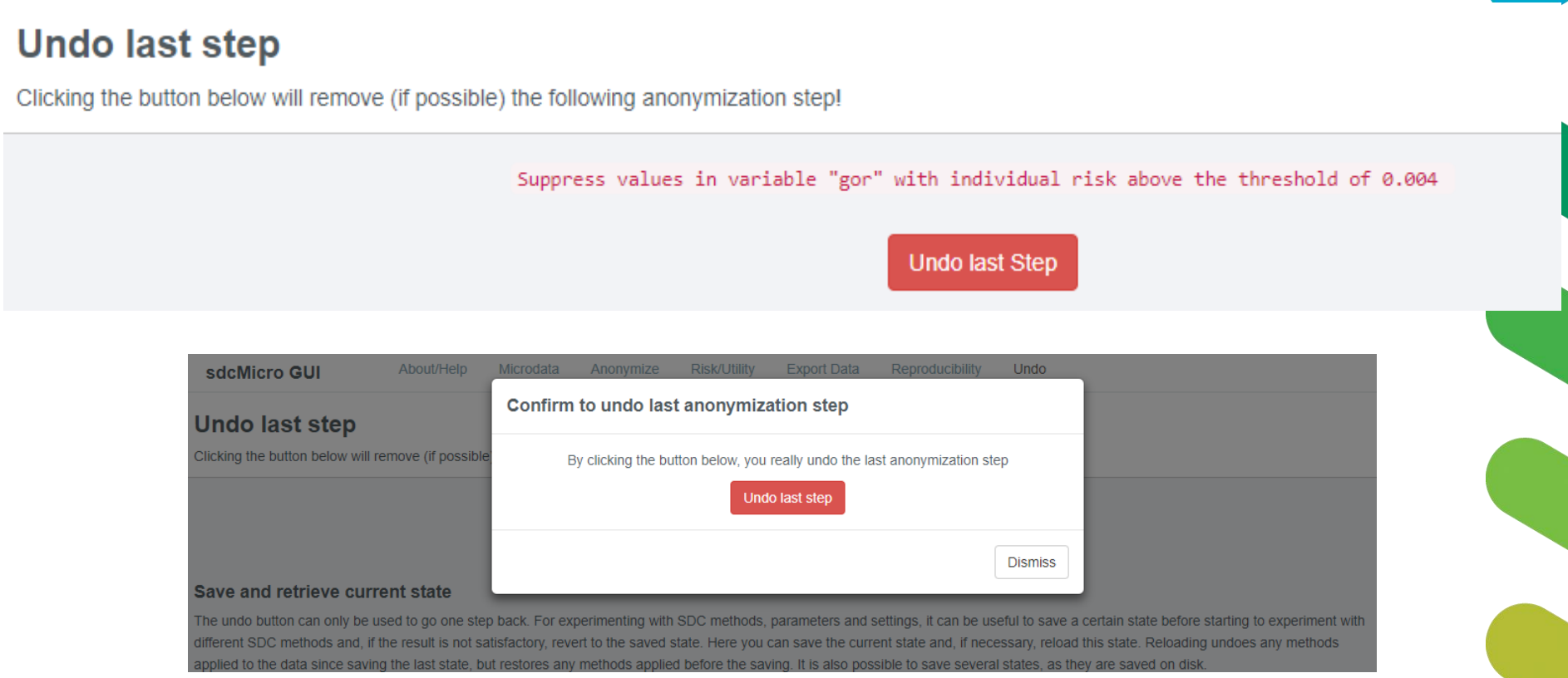
2nd Step: K-Anonymization
0.91% to 0.13%

3rd Step: Suppression of values with high risk
0.13% to 0.13%

Was the 3rd step worth it?
NO



As 3rd step did not make a significant difference we can undo the step



Caution: You can only undo last step

UK Data Service

View/Analyze existing sdcProblem

Show summary

Explore variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PRAM (simple)

PRAM (expert)

Supress values with high risks

Anonymize numerical variables

Top/bottom coding

Microaggregation

Adding noise

Rank swapping

Risk measures for categorical key variables

We expect `4.79` (`0.13%`) re-identifications in the population, as compared to `44.14` (`1.19%`) re-identifications in the original data.

`0` observations have a higher risk than the risk in the main part of the data, as compared to `0` observations in the original data.

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	1059 (28.514%)
3-anonymity	0 (0.000%)	1659 (44.669%)
5-anonymity	590 (15.886%)	2429 (65.401%)

Information on local suppression

Key variable	Additional suppressions due to last run of kAnon()	Total number of missing values (NA) in variable
rsex	2 (0.054%)	2 (0.054%)
rethnic	10 (0.269%)	10 (0.269%)
relig	643 (17.313%)	643 (17.313%)
highqual	51 (1.373%)	51 (1.373%)
occup	545 (14.674%)	545 (14.674%)
gor	127 (3.419%)	127 (3.419%)

Happy with the anonymized data? If yes it can be exported in 5 different formats

What do you want to export?

- Anonymized Data
- Anonymization Report

Export anonymized microdata

Select the file format to export the data to. If necessary, the order of the records can be randomized before exporting.

View anonymized data

Show 10 entries Search:

rsex	rage	rethnic	relig	highqual	occup	cancer	car	weight	gor	frigwork
Male	18	Black	Christianity	GCSE (D-E)		No	No	135	NI	Yes
Male	18	Black	Islam	Undergraduate	Ambulance Officer	Yes	Yes	126	NI	Yes
Male	18	Black		GCSE (D-E)	Baker	No	No	112	London	Yes
Male	18	Black		A-levels	Building Inspector	Yes	Yes	143	London	Yes
Female	18	Black	No religion		Cardiologist	No	No	199	London	Yes
Female	18	Black	Hinduism	Postgraduate - PHD	Clinical Psychologist	Yes	Yes	185		Yes
Female	18	White	Christianity	GCSE (A-C)	Cook	No	No	245	London	Yes
Female	18	Black	Christianity	GCSE (A-C)	Electrical Engineer	Yes	Yes	202	London	Yes
Male	18	Black	Christianity	Postgraduate - MA, MSc		No	No	101	London	Yes
Male	18	Black	Christianity	High School	Landscape Gardener	Yes	Yes	193	London	Yes

Showing 1 to 10 of 3,714 entries

Previous 1 2 3 4 5 … 372 Next

Select file format for export

(•) R-dataset (.RData) ( ) SPSS-file (.sav) ( ) CSV-file (.csv) ( ) STATA-file (.dta) ( ) SAS-file (.sas7bdat)

You can also obtain an **anonymization report** (it will be **saved in your set storage path**)

What do you want to export?

Anonymized Data

Anonymization Report

Create anonymization report

A report for internal use (more detailed) or a report for external use (less detailed) is saved to the export directory.

Select type of report

(•) internal (detailed) ( ) external (short overview)

Save report

SDC-Report

// Input Data

The data set consists of **3714** observations and was imported from **test_data.sav**.

// Information on selected important (key) variables

- **Categorical key variable(s)**: *rsex | rethnic | relig | highqual | occup | gor*
- **Continuous key variable(s)**: *rage*
- **Weight variable**: *weight*
- **householdID**: *not defined*
- **strataVariable(s)**: *not defined*

// Modifications

- Modifications on categorical key variables: **TRUE**
- Modifications on continuous key variables: **FALSE**
- Modifications using PRAM: **FALSE**
- Local suppressions: **TRUE**

// Disclosure risk:

/// Frequency Analysis for Categorical Key Variables

//// Number of observations violating

- **2-Anonymity:** 0 (original dataset: 1059)
- **3-Anonymity:** 0 (original dataset: 1659)

//// Percentage of observations violating

- **2-Anonymity:** 0.000% (original dataset: 28.514%)
- **3-Anonymity:** 0.000% (original dataset: 44.669%)

/// Disclosure Risk for Categorical Variables

Expected Percentage of Reidentifications:

- **modified data**: 0.129% (~ 4.790 observations)
- **original data**: 1.189% (~ 44.142 observations)

//// 10 combinations of categories with highest risks

rsex	rethnic	relig	highqual	occup	gor	risk	fk	Fk
Male	Asian	Christianity	High School	NA	NI	0.005	3	317
Male	Mixed	No religion	Undergraduate	Accountant	NA	0.005	3	318
Male	Mixed	NA	Undergraduate	Accountant	West Midlands	0.005	3	318

And lastly, you can save the script use to create the anonymized data file for reproducibility

What do you want to do?

View the current script

Import a previously saved problem

Export/Save the current sdcProblem

View the current generated script

Browse and download the script used to generate your results. These can be used later as a reminder of what you did or entered into R from command-line to reproduce results.

Save Script to File

```
require(sdcMicro)
inputdata <- readMicrodata(path="C:/Users/dcmagd/AppData/Local/Temp/RtmpOECAHN/842d8aefee7269281cc3d5eb/test_data.sav", type="spss", c
onvertCharToFac=TRUE, drop_all_missings=TRUE)
inputdataB <- inputdata

## Set up sdcMicro object
sdcObj <- createSdcObj(dat=inputdata,
        keyVars=c("rsex","rethnic","relig","highqual","occup","gor"),
        numVars=c("rage"),
        weightVar=c("weight"),
        hhId=NULL,
        strataVar=NULL,
        pramVars=NULL,
        excludeVars=NULL,
        seed=0,
        randomizeRecords=FALSE,
        alpha=c(1))
```

UK Data Service

Nota Bene

- Do not include more than 12 key variables in an SDC problem (**ideally 5-7**); can test for several SDC problems (different key variables combinations)
- Try **not to use more than 2 different anonymization techniques**
- Always think at **data utility** and take into account the **information loss**
- **Risk is highly dependable on the weighting factor** (if no weight is available risk will be significantly higher see slide 30)
- Any SDC tools are useful for **providing comparison** between SDC methods and sdcMicro GUI provides a **quick and easy** way to explore what **changes have biggest effect** – however - **the numbers might have no real meaning** –
- **There is no absolute 0% risk**

No weighting used

Summary of dataset and variable selection

The loaded dataset consists of `3714` records and `12` variables.

Categorical key variable(s): `rsex` `rethnic` `relig` `highqual` `occup` `gor`
Numerical key variable(s): `rage`

Computation time

The current computation time was ~ `30.04 seconds` .

Same data and same model used but without weighting

Risk measures

The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

(•) Risk measures ( ) Risky observations ( ) Plot of risk

Risk measures

`1059` observations have a higher risk than the risk in the main part of the data, as compared to `1059` observations in the original data ⓘ

Based on the individual re-identification risk, we expect `1770` re-identifications (`47.66%`) in the anonymized data set. In the original dataset we expected `1770` (`47.66%`) re-identifications.

Without weighting risk is 47.66% compared to **with weighting when risk was 1.19%**

Always be careful when interpreting risk

UK Data Service

Resources

- **sdcMicro Reference Manual**
- **sdcMicro GUI Vignette**
- **Guidelines for statistical disclosure control using sdcMicro Vignette**
- **sdcMicro Git Page**
- Matthias Templ, Alexander Kowarik, Bernhard Meindl (2015). *Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro*. Journal of Statistical Software, 67(4), 1-36. doi:10.18637/jss.v067.i04

UK Data Service

Questions

Cristina Magder

dcmagd@essex.ac.uk