Organisation des Nations Unies
pour l'alimentation et l'agriculture

Opening Access to Agricultural Microdata

3 – Statistical Disclosure Control (SDC)

Regional webinar - Africa
10:00 to 13:00 (GMT+0), November 29, 2021

Outline

What is disclosure risk and Statistical Disclosure Control (SDC)?

- Key Steps in SDC process
- SDC outputs

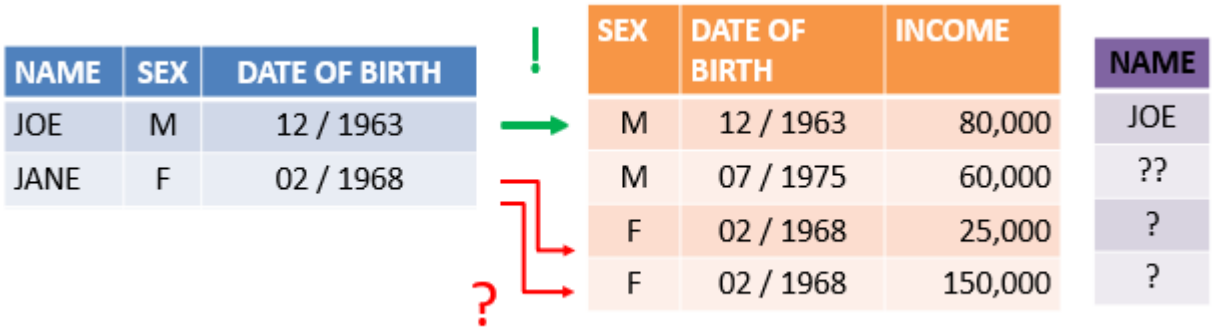
NB: presentation will be accompanied with screenshots from sdcApp(), a user interface for microdata anonymization in R with the sdcMicro Package

What is disclosure risk?

Disclosure occurs when a person or an organization (intruder) recognizes or learns something that they did not already know about another person or organization through released data
(Source: Australian Bureau of Statistics)

How do intruders re-identify respondents?

Assume intruder has a data file with names and key variables, allowing record linkage



Types of Disclosure

Identity disclosure : occurs if an intruder associates a known individual with a released data record (Lambert, 1993)

Attribute disclosure : occurs if the intruder is able to determine some new characteristics of an individual based on the information available in the released data (Lambert, 1993)

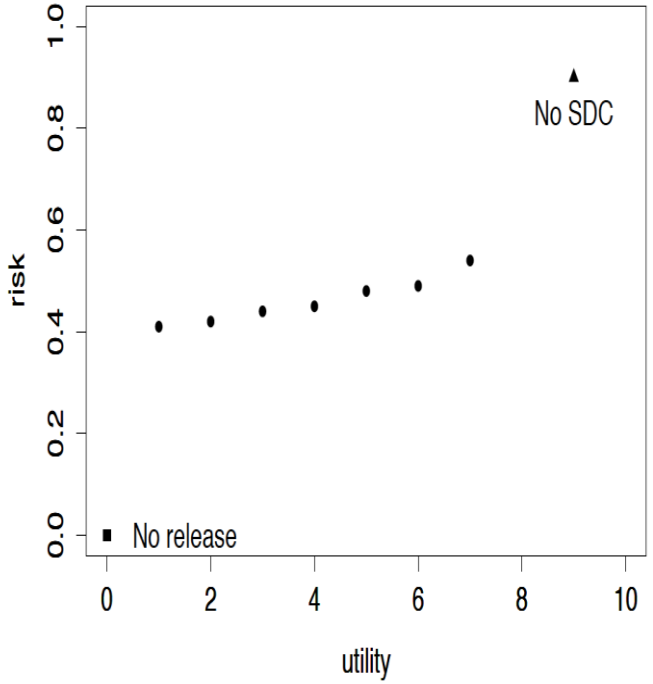
Inferential disclosure : occurs when the intruder is able to determine the value of some characteristic of an individual more accurately with the released data than otherwise would have been possible (Lambert, 1993)

What is SDC?

- SDC is a process of treating data in order to reduce the risk of disclosure
- The aim being to achieve an “acceptable level” of disclosure risk and allow for appropriate release
- The level of acceptability of disclosure risk is usually at the discretion of the data producer and guided by legislation as well as by ethical concerns

Risk-Utility Trade-off

- Treating data to reduce disclosure risk always reduces the utility of the data to end users
- The goal is to produce a dataset with low risk of identification and maximum utility
- All SDC methods result in some level of information loss
- SDC will reduce risk, but risk is never eliminated
- In general the greater the level of protection the higher the information loss and loss of utility



Applying SDC: Stepwise approach

- The SDC process can be divided in several steps
- Some steps are iterative
- Step-wise approach as guideline, not definitive order

Overview of steps

- Step 0 : Need for confidentiality protection
- Step 1 : Data preparation and exploring data characteristics
- Step 2: Type of release
- Step 3 : Intruder scenarios and choice of key variables
- Step 4 : Data key uses and selection of utility measures
- Step 5 : Assessing disclosure risk
- Step 6 : Assessing utility measures
- Step 7 : Choice and application of SDC methods
- Step 8 : Re-measure risk
- Step 9 : Re-measure utility
- Step 10 : Audit and reporting
- Step 11 : Data release

Types of SDC and existing tools

In this presentation - SDC applied in microdata

However, main theoretical principles of SDC concerning

- Tabular data
- Microdata protection
- Output checking

A broad scientific literature on SDC has been established and SDC is practiced in many NSOs and other microdata producers around the world.

Tools:

- μ-ARGUS (https://research.cbs.nl/casc/mu.htm)
- sdcMicro (in R Statistical software)

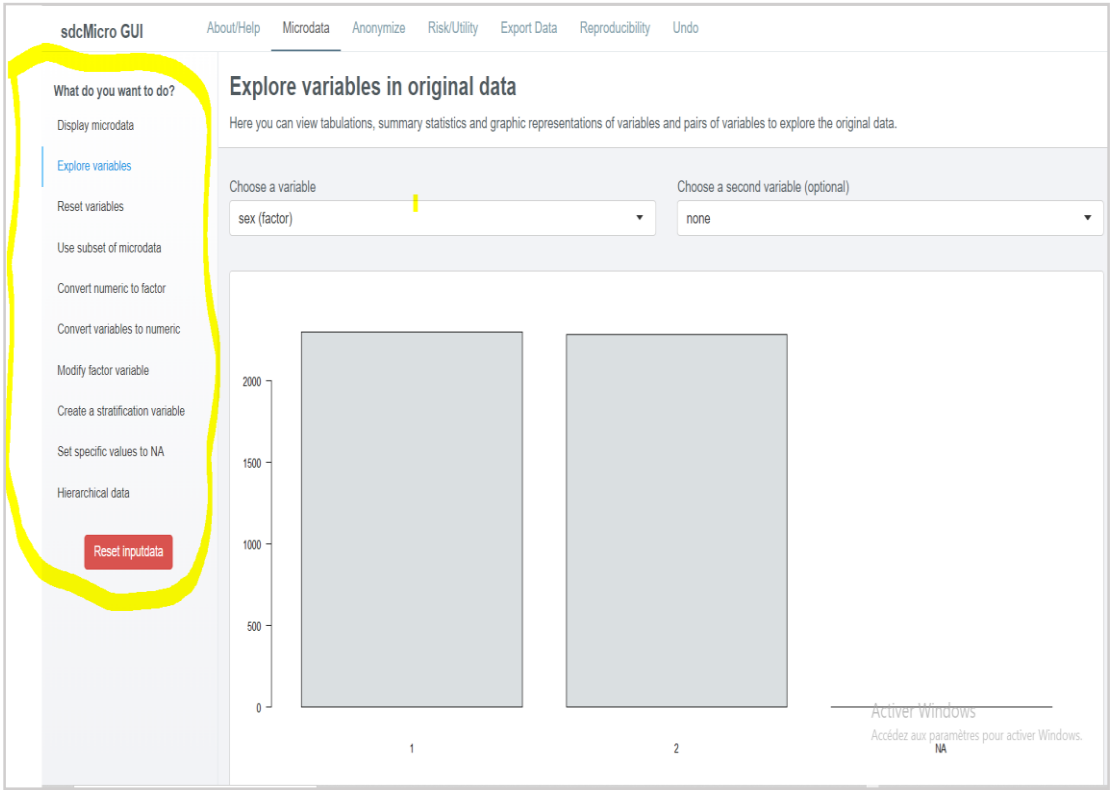
Step 0: Need for confidentiality protection

- Need for confidentiality has to be determined
- Interpretation of laws and regulations
- Determine what the statistical units at risk are
- If individuals, households, legal entities likely then need for disclosure control

Step 1: Data preparation and exploring data characteristics

Exploring data

- What are the variables? Are there many missing values?
- What are the main uses of the data?
- Classify variables as sensitive and non-sensitive
- Remove direct identifiers
- Variables with many missing values are removed
- Information about survey methodology (strata, sampling weights, relationships between variables)

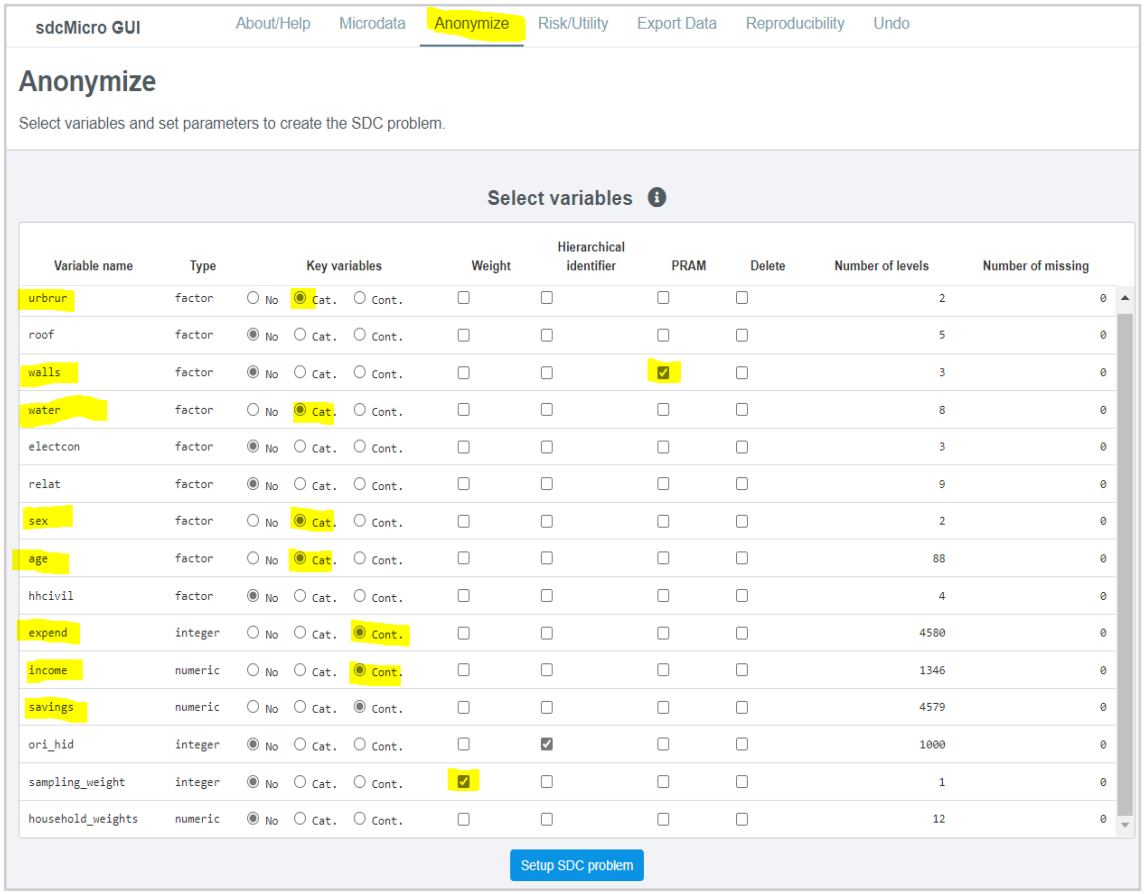


Step 2: Type of release

- Choice of release type
  - Possible multiple release types
  - All information of less detailed release should be contained in more detailed release
  - For one release type only a single release (no releases tailored to user)
- Determines the required level of protection
- Some data cannot be released: too sensitive or not possible to protect sufficiently

Step 3: Intruder scenarios and choice of key variables

- Defining disclosure scenarios
  - Type of external data sources available
  - Ways the intruder uses these data sources
- Determining key variables
- This has to be done for each release type

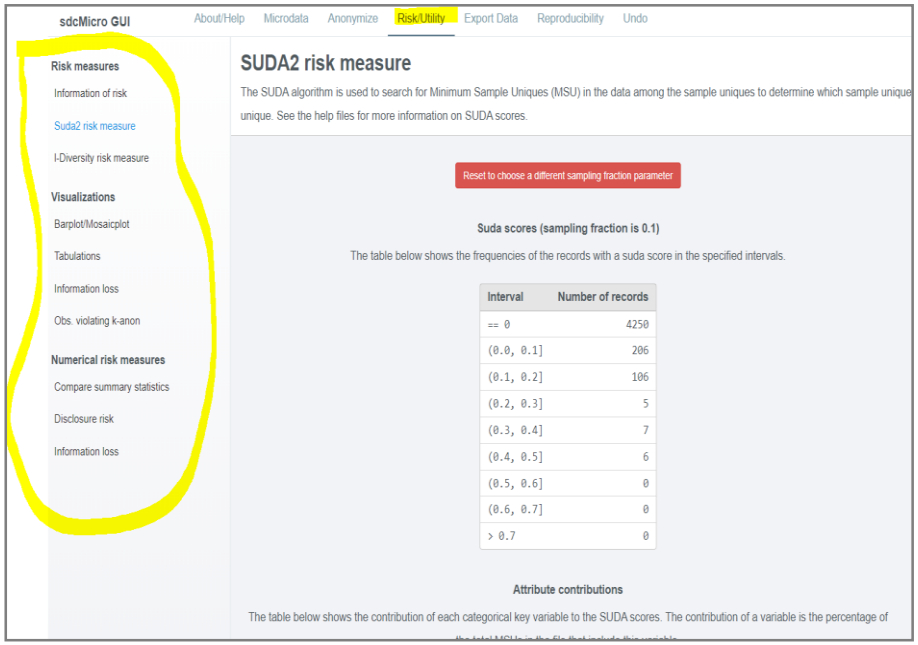


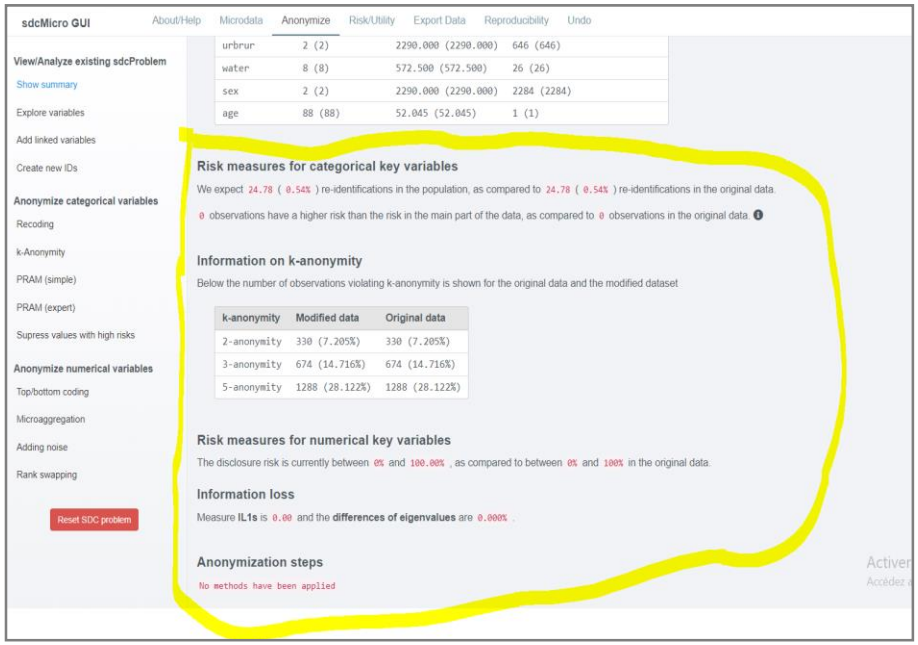
Step 4: Data key uses and selection of utility measures

- Determining key uses of the data (most common statistical analyses)
- Pick utility measures for the dataset based on these key uses

Step 5: Assessing disclosure risk

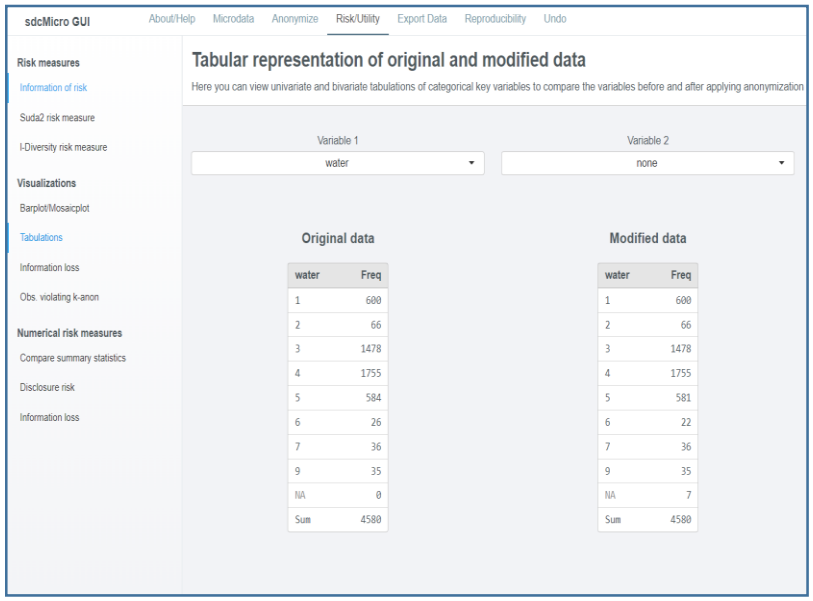
- Select appropriate risk measures
- Evaluate risk

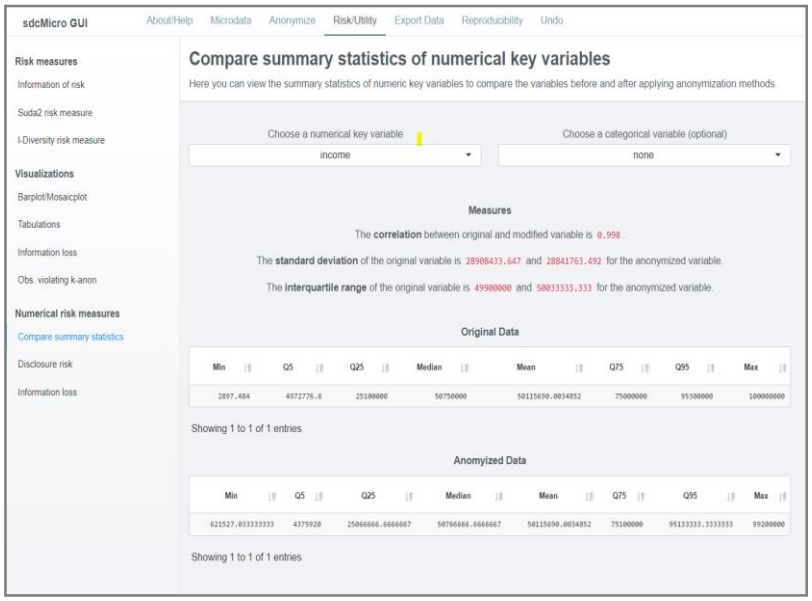




Step 6: Assessing utility measures

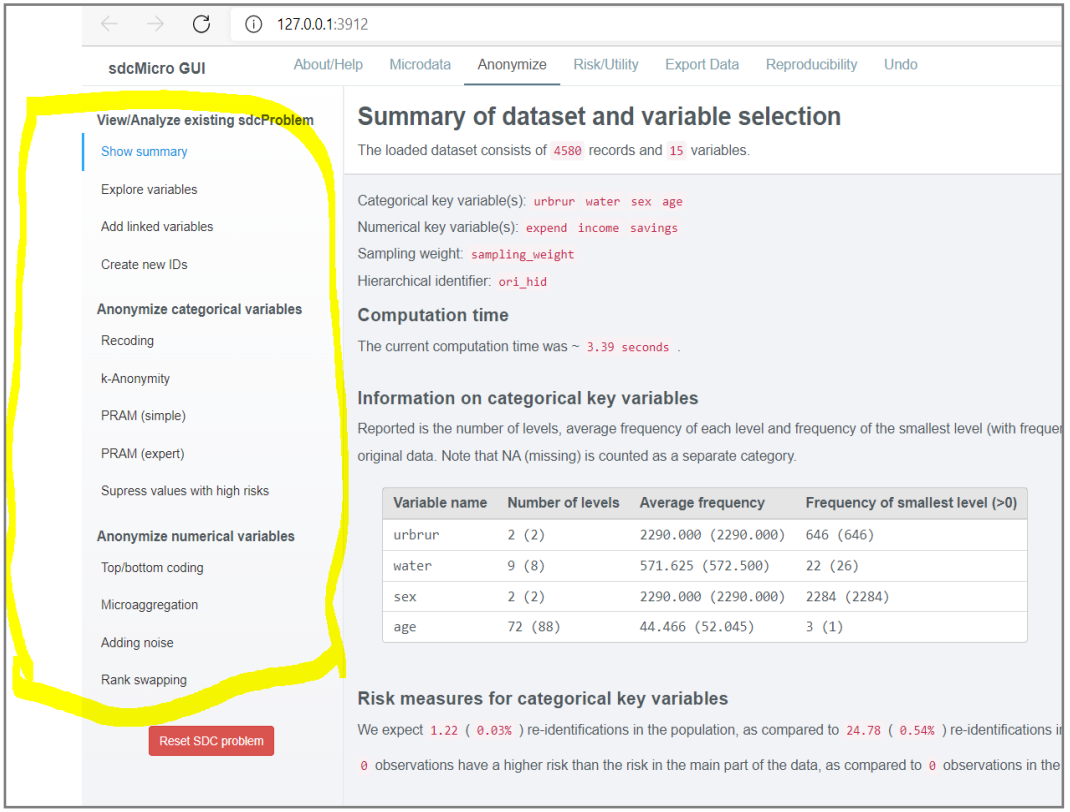
- Evaluate the utility measures selected
- Create a set of benchmark values for comparison after anonymization





Step 7: Choice and application of SDC methods

- Choice of SDC methods
- Iterative approach (trial and error)
- After applying each method risk and utility are re-measured and compared with the initial values
- Searching for appropriate methods and parameters (reiterative, trial-and-error)



Step 8 : Re-measure risk

- Re-measure risks
- Special attention to outliers

Step 9: Re-measure utility

- Re-measure utility
- Compare with initial values

Note: Hierarchical structure

- Where there is a hierarchical structure (e.g., household structure) in the data, step 5 through 9 should be first done with only the households variables and after that repeated with both household and individual variables

Step 10: Audit and reporting

- Check on relationships between variables, consistency, unusual values
- Reporting
  - Internal for quality purposes \ supervisory authorities
  - External for data users: data users know for what purposes the data is valid
  - Information loss should be explained to users
  - More generally policies relating to what level of information is published should be made available on your website: level of, geography, occupation codes etc. per access type

Step 11: Data release

- Actual data release
- Under appropriate access type for the level of treatment, risk assessment, sensitivity

SDC report

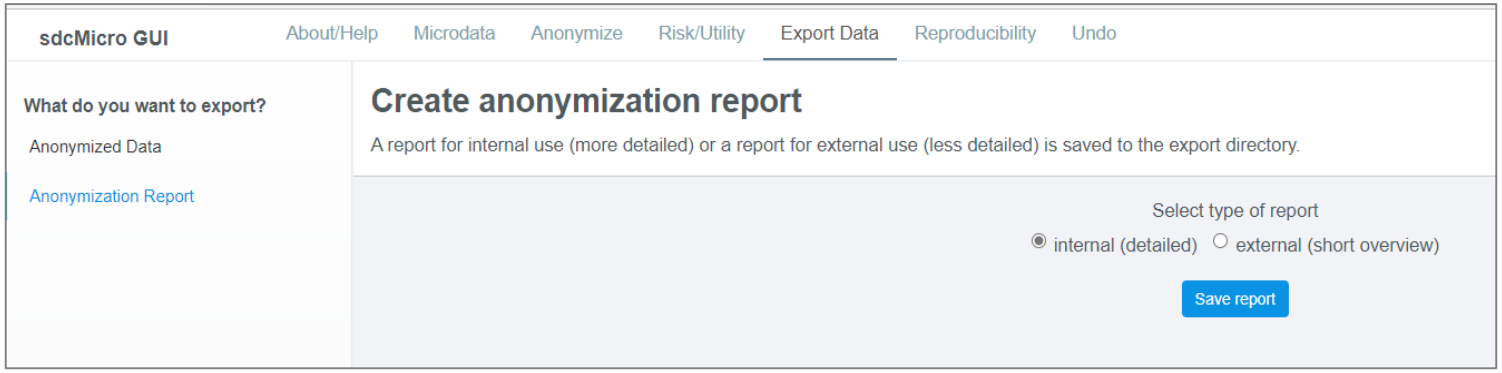
Outputs of SDC:

1. Anonymized microdata

- It is the dataset ready for release

2. An SDC report

- The SDC report is very important as it allows to keep track of the different methods that are apply through the years.
- Important for the institution memories
- R offers a possibility to generate report that incorporate the anonymization results from the sdcMicro package (Rmarkdown).



Thank you

SDC – Useful resources:

- Thijs Benschop and Matthew Welch, "*Statistical Disclosure Control for Microdata: A Practice Guide*" (June 2016), https://sdcpractice.readthedocs.io/en/latest/

- Alexander Kowarik, Bernhard Meind, and Matthias Templ, "*Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro*" file:///C:/Users/vanderm/Downloads/v67i04%20(1).pdf

- https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf