



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

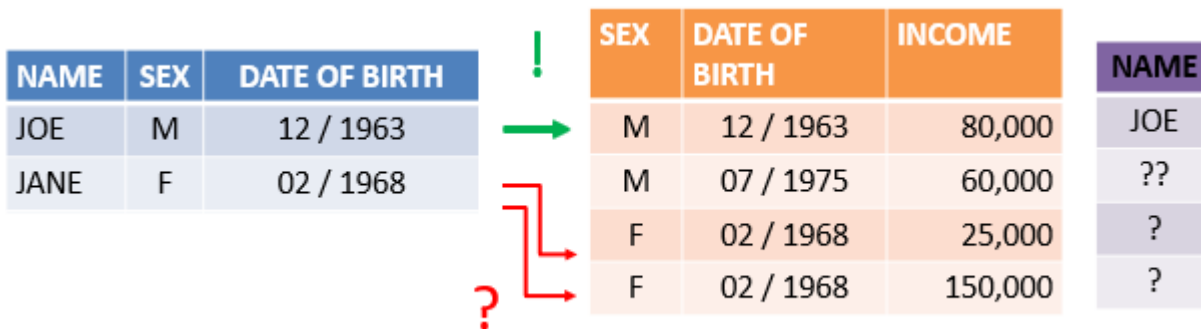
NB: presentation will be accompanied with screenshots from `sdcApp()`, a user interface for microdata anonymization in R with the `sdcMicro` Package

# 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)

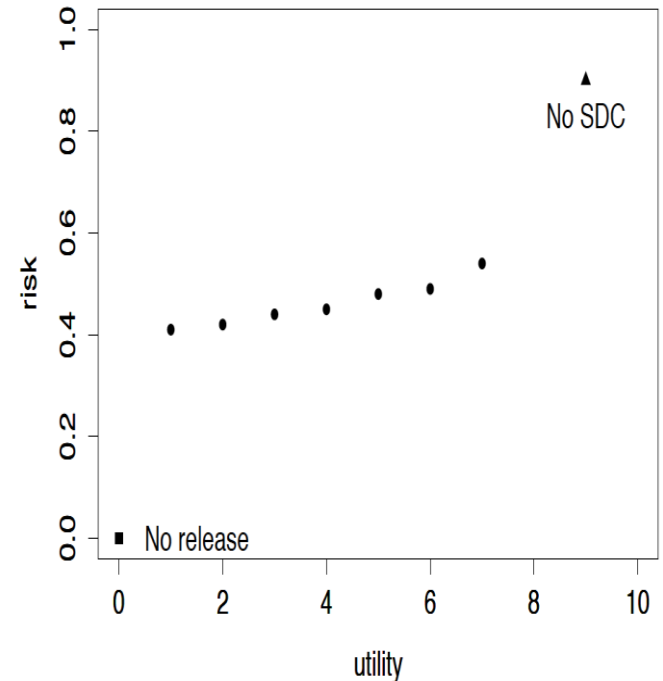
**Inferential disclosure** : occurs when the intruder is able to determine the value of some characteristic of an individual more accurately with the released data than otherwise would have been possible (Lambert, 1993)

# 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:

- $\mu$ -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)

sdcMicro GUI

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

### Explore variables in original data

Here you can view tabulations, summary statistics and graphic representations of variables and pairs of variables to explore the original data.

Choose a variable: sex (factor)

Choose a second variable (optional): none

What do you want to do?

- Display microdata
- Explore variables
- Reset variables
- Use subset of microdata
- Convert numeric to factor
- Convert variables to numeric
- Modify factor variable
- Create a stratification variable
- Set specific values to NA
- Hierarchical data

Reset input data

2000  
1500  
1000  
500  
0

1 2

Activer Windows  
Accédez aux paramètres pour activer Windows.  
NA

## 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

sdcMicro GUI    About/Help    Microdata    **Anonymize**    Risk/Utility    Export Data    Reproducibility    Undo

## 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
urbrur	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2	0
roof	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	5	0
walls	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	3	0
water	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8	0
electcon	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	3	0
relat	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	9	0
sex	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2	0
age	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	88	0
hhcivil	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4	0
expend	integer	<input type="radio"/> No	<input type="radio"/> Cat.	<input checked="" type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4580	0
income	numeric	<input type="radio"/> No	<input type="radio"/> Cat.	<input checked="" type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1346	0
savings	numeric	<input type="radio"/> No	<input type="radio"/> Cat.	<input checked="" type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4579	0
ori_hid	integer	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1000	0
sampling_weight	integer	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1	0
household_weights	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	12	0

Setup SDC problem

## **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

**sdcmicro GUI** About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

### 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 unique. See the help files for 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	4250
(0.0, 0.1]	206
(0.1, 0.2]	106
(0.2, 0.3]	5
(0.3, 0.4]	7
(0.4, 0.5]	6
(0.5, 0.6]	0
(0.6, 0.7]	0
> 0.7	0

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

**sdcmicro GUI** About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

Variable	Count	Original Count	Original Percentage
urbrur	2 (2)	2290.000 (2290.000)	646 (646)
water	8 (8)	572.500 (572.500)	26 (26)
sex	2 (2)	2290.000 (2290.000)	2284 (2284)
age	88 (88)	52.045 (52.045)	1 (1)

### Risk measures for categorical key variables

We expect 24.78 ( 0.54% ) re-identifications in the population, as compared to 24.78 ( 0.54% ) 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	330 (7.205%)	330 (7.205%)
3-anonymity	674 (14.716%)	674 (14.716%)
5-anonymity	1288 (28.122%)	1288 (28.122%)

### Risk measures for numerical key variables

The disclosure risk is currently between 0% and 100.00% , as compared to between 0% and 100% in the original data.

### Information loss

Measure IL1s is 0.00 and the differences of eigenvalues are 0.000% .

### Anonymization steps

No methods have been applied

# Step 6: Assessing utility measures

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

The screenshot shows the 'sdcmicro GUI' interface. The main panel is titled 'Tabular representation of original and modified data'. It includes a sidebar with navigation options like 'Risk measures', 'Visualizations', and 'Tabulations'. The main content area has two dropdown menus for 'Variable 1' (set to 'water') and 'Variable 2' (set to 'none'). Below these are two tables: 'Original data' and 'Modified data'. Both tables show the frequency distribution for the 'water' variable across categories 1 to 9, with a total sum of 4580.

Original data		Modified data	
water	Freq	water	Freq
1	600	1	600
2	66	2	66
3	1478	3	1478
4	1755	4	1755
5	584	5	581
6	26	6	22
7	36	7	36
9	35	9	35
NA	0	NA	7
Sum	4580	Sum	4580

The screenshot shows the 'sdcmicro GUI' interface. The main panel is titled 'Compare summary statistics of numerical key variables'. It includes a sidebar with navigation options like 'Risk measures', 'Visualizations', and 'Numerical risk measures'. The main content area has two dropdown menus for 'Choose a numerical key variable' (set to 'income') and 'Choose a categorical variable (optional)' (set to 'none'). Below these are several text-based statistics: 'The correlation between original and modified variable is 0.998', 'The standard deviation of the original variable is 28908433.647 and 28841763.492 for the anonymized variable.', and 'The interquartile range of the original variable is 49980000 and 50033333.333 for the anonymized variable.' There are also two summary statistics tables: 'Original Data' and 'Anonymized Data', both showing Min, Q5, Q25, Median, Mean, Q75, Q95, and Max values.

Original Data							
Min	Q5	Q25	Median	Mean	Q75	Q95	Max
2897.484	4972776.6	25100000	50750000	50115090.0034852	75000000	95300000	100000000

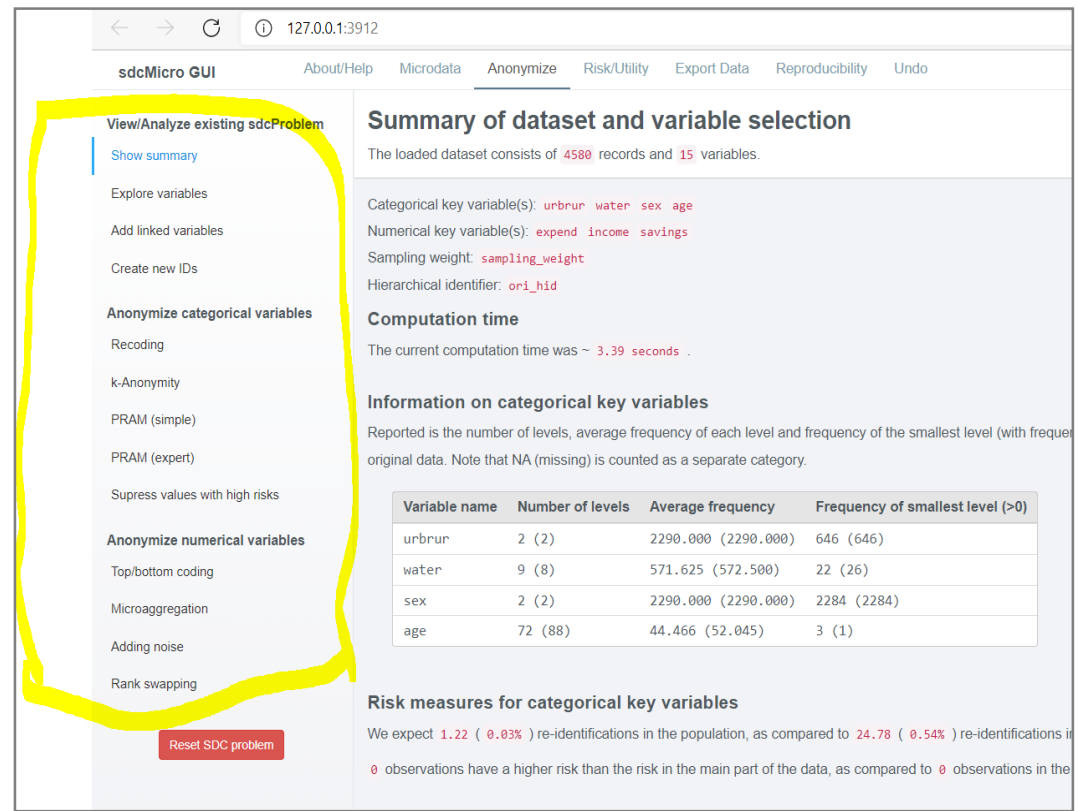
  

Anonymized Data							
Min	Q5	Q25	Median	Mean	Q75	Q95	Max
623527.833333333	4379920	25866666.6666667	50766666.6666667	58115090.0034852	75100000	95133333.3333333	95200000



# 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)



The screenshot displays the sdcMicro GUI interface. The left sidebar contains a menu of options, with the top section highlighted in yellow. The main panel shows a summary of the dataset and variable selection, including a table of categorical key variables and risk measures.

**sdcMicro GUI** About/Help Microdata Anonymize Risk/Utility Export Data Reproducibility Undo

**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

**Summary of dataset and variable selection**

The loaded dataset consists of 4580 records and 15 variables.

Categorical key variable(s): urbrur water sex age  
Numerical key variable(s): expend income savings  
Sampling weight: sampling\_weight  
Hierarchical identifier: ori\_hid

**Computation time**

The current computation time was ~ 3.39 seconds .

**Information on categorical key variables**

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency of the smallest level). Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
urbrur	2 (2)	2290.000 (2290.000)	646 (646)
water	9 (8)	571.625 (572.500)	22 (26)
sex	2 (2)	2290.000 (2290.000)	2284 (2284)
age	72 (88)	44.466 (52.045)	3 (1)

**Risk measures for categorical key variables**

We expect 1.22 ( 0.03% ) re-identifications in the population, as compared to 24.78 ( 0.54% ) re-identifications in the population. 0 observations have a higher risk than the risk in the main part of the data, as compared to 0 observations in the population.

Reset SDC problem

## 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

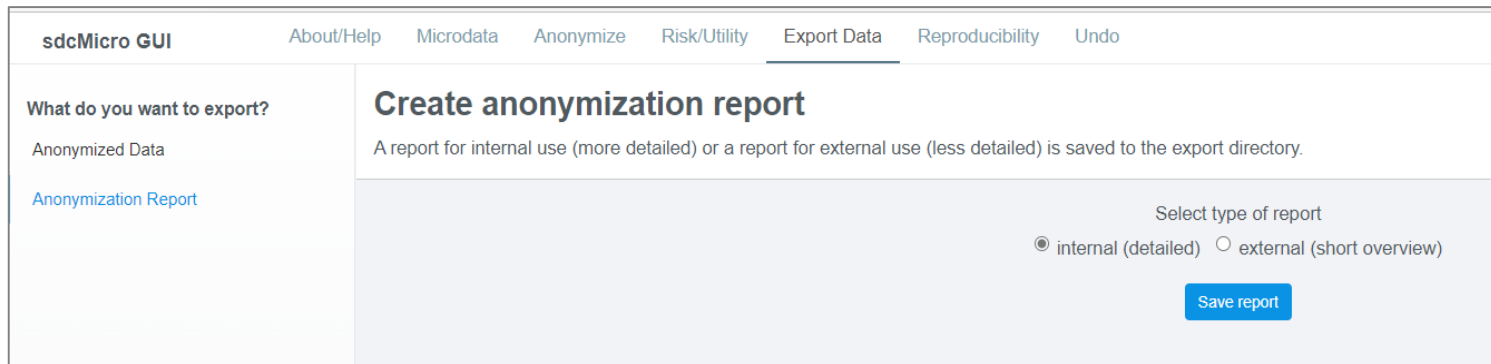
## Outputs of SDC:

### 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).



The screenshot shows the sdcMicro GUI interface. The top navigation bar includes 'sdcMicro GUI', 'About/Help', 'Microdata', 'Anonymize', 'Risk/Utility', 'Export Data' (which is the active tab), 'Reproducibility', and 'Undo'. On the left side, under 'What do you want to export?', there are two options: 'Anonymized Data' and 'Anonymization Report' (which is selected). The main content area is titled 'Create anonymization report' and contains the text: 'A report for internal use (more detailed) or a report for external use (less detailed) is saved to the export directory.' Below this text, there is a section for 'Select type of report' with two radio button options: 'internal (detailed)' (which is selected) and 'external (short overview)'. At the bottom right of this section is a blue 'Save report' button.

# 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](file:///C:/Users/vanderm/Downloads/v67i04%20(1).pdf)
- [https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc\\_guidelines.pdf](https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf)