Anonymizing data using SDC

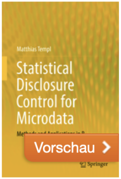
Matthias Templ

Institut für Datenanalyse und Prozessdesign
School of Engineering
Zürcher Hochschule für Angewandte Wissenschaften

BernR User Group R, 11.11.2021

Background in Anonymization

- Projects in/with Statistics Austria, OECD, IHSN, Weltbank, EU, Helsana, Swisscom, Malawi, SBB, Stadt and Canton of Zurich, Consultations for BAG, Publications, Workshops, . . .
- Springer-Book Statistical Disclosure Control



© 2017

Statistical Disclosure Control for Microdata

Methods and Applications in R

Autoren: **Templ**, Matthias

- Lecture *Advanced Survey Statistics: Statistical Disclosure Control* at the free Univ. of Berlin, Bamberg and Trier (2019, 2020).

Typical problems

Everywhere when detailed data sets containing individual personal information needs to be shared

In Business

- Companies store and distribute (internally or externally) data that includes customer information
- ...

Health insurance

- shares detailed data with universities and hospitals for analysis purposes
- Example: Sharing health data

Typical problems

Everywhere when detailed data sets containing individual personal information needs to be shared

In Business

- Companies store and distribute (internally or externally) data that includes customer information
- ...

Health insurance

- shares detailed data with universities and hospitals for analysis purposes
- Example: Sharing health data

Open or/and scientific data in official statistics

- open data becomes important, and there is an increasing need to share data
- scientific use: researchers who need detailed data
- Example: Sharing health data

Overview:

- Important types of characteristics for anonymisation
- Quantifying the disclosure risk
- Anonymisation of data
- Quality assessment of the anonymised data

IT security is not discussed. It does, however, play a role with regard to the degree of anonymisation required.

- The smaller the IT security, the more rigid anonymisation necessary

Overview:

- Important types of characteristics for anonymisation
- Quantifying the disclosure risk
- Anonymisation of data
- Quality assessment of the anonymised data

IT security is not discussed. It does, however, play a role with regard to the degree of anonymisation required.

- The smaller the IT security, the more rigid anonymisation necessary

Anonymisation of socio-demographic characteristics is very central - this will be the first (and the only) longer part.

Also **aggregated information** (tabular data) is worth protecting.

People may also be identified by movement patterns → **anonymisation of trajectory data**.

−> −> −>

What to do?

- (ISO/TS 25237:2008) Anonymization: *Process that removes the association between the identifying data set and the data subject.*
- Anonymisation involves the **use of complex methods** of statistical disclosure control.
- *Absolute anonymity* is not possible and is not required by e.g. the DSGVO or Swiss DSG (keyword **de-facto anonymity**)

What to do?

- (ISO/TS 25237:2008) Anonymization: *Process that removes the association between the identifying data set and the data subject.*
- Anonymisation involves the **use of complex methods** of statistical disclosure control.
- *Absolute anonymity* is not possible and is not required by e.g. the DSGVO or Swiss DSG (keyword **de-facto anonymity**)

de-facto anonymity

If the **effort is higher** data is to be re-identified **as the benefit** we speak of **de-facto anonymity**.

Anonymisation in practice: Rough procedure

1) **RISK** Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called *disclosure scenario*.
- Individual risk (of each individual person) and global risk

Anonymisation in practice: Rough procedure

1) RISK Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called *disclosure scenario*.
- Individual risk (of each individual person) and global risk

2) Anonymisation 

- Traditional methods or synthetic data generation?
- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?

1) RISK Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called *disclosure scenario*.
- Individual risk (of each individual person) and global risk

2) 

Anonymisation 

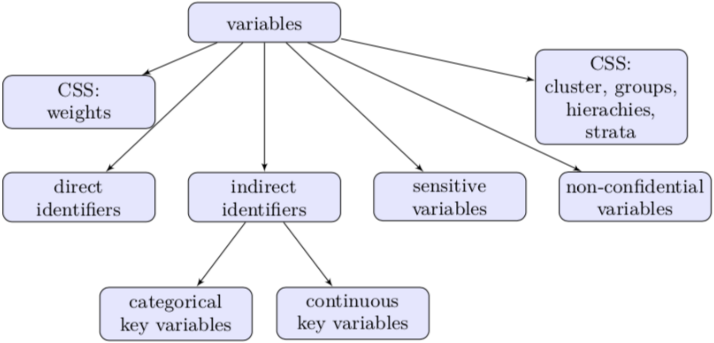
- Traditional methods or synthetic data generation?
- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?

3) 

Measurement of the utility 

- Global procedures or data-specific comparisons?
- What is the analysis of interest of the users?

Variable types



We will take a closer look at these distinctions below...

Types of characteristics in data (1/2)

1. **delete** globally unique (e.g. insurance number) and **direct identifiers** (e.g. exact address or name) or **pseudo-anonymise** (with area and project-specific salts and hashes)

Types of characteristics in data (1/2)

1. **delete** globally unique (e.g. insurance number) and **direct identifiers** (e.g. exact address or name) or **pseudo-anonymise** (with area and project-specific salts and hashes)
2. **Quasi-identifiers** (e.g. postcode, age, gender), abbreviated QIDs: Attributes that can be used for re-identification; are called also **Key Variables**, *Indirect Identifiers* or *Implicit Identifiers*. Slopp: Those variables that overlap with other populations (or samples) available on the market.

Types of characteristics in data (1/2)

1. **delete** globally unique (e.g. insurance number) and **direct identifiers** (e.g. exact address or name) or **pseudo-anonymise** (with area and project-specific salts and hashes)
2. **Quasi-identifiers** (e.g. postcode, age, gender), abbreviated QIDs: Attributes that can be used for re-identification; are called also **Key Variables**, *Indirect Identifiers* or *Implicit Identifiers*. Slopp: Those variables that overlap with other populations (or samples) available on the market.

A **Key** defines a combination of QID's (e.g. age $=$ 10, gender $=$ M, region $=$ ZH)

Example: Matching of key variables:

Records released (QID's: residence, occupation, gender)

name	place of residence	profession	sex	# ...	Default	Income
x	Stadel	Prof	M	1	...	yes
x	Winterthur	architect	M	18	...	no
x	...	...	...	...	...	...

External data set from GfK

name	place of residence	profession	gender	#
Max Muster	Stadel	Prof	M	1
Jo Johann	Winterthur	Architect	M	18
Nils Nilson	Winterthur	Architect	M	18
...	...	...	...	...

Example: Matching of key variables:

Records released (QID's: residence, occupation, gender)

name	place of residence	profession	sex	# . . .	Default	Income
x	Stadel	Prof	M	1	. . .	yes
x	Winterthur	architect	M	18	. . .	no
x	. . .	. . .	. . .	. . .	. . .	. . .

External data set from GfK

name	place of residence	profession	gender	#
Max Muster	Stadel	Prof	M	1
Jo Johann	Winterthur	Architect	M	18
Nils Nilson	Winterthur	Architect	M	18
. . .	. . .	. . .	. . .	. . .

$\rightarrow$ Max Muster is clearly matchable.

Types of characteristics in data (2/2)

3. **Sensitive attributes** (e.g. sickness status, costs, late payments, mental disorder, ...): Information that individuals do not want to be associated with.

Types of characteristics in data (2/2)

3. **Sensitive attributes** (e.g. sickness status, costs, late payments, mental disorder, ...): Information that individuals do not want to be associated with.
4. **Linked** (also sometimes called *Ghost* variables), example: If region is suppressed, the information on municipality must also be suppressed.

Types of characteristics in data (2/2)

3. **Sensitive attributes** (e.g. sickness status, costs, late payments, mental disorder, . . .): Information that individuals do not want to be associated with.
4. **Linked** (also sometimes called *Ghost* variables), example: If region is suppressed, the information on municipality must also be suppressed.
5. **weight vector**: If available. Sample weights have to be considered for SDC. Risk of identification is higher for population data than for sample data.

Types of characteristics in data (2/2)

3. **Sensitive attributes** (e.g. sickness status, costs, late payments, mental disorder, ...): Information that individuals do not want to be associated with.
4. **Linked** (also sometimes called *Ghost* variables), example: If region is suppressed, the information on municipality must also be suppressed.
5. **weight vector**: If available. Sample weights have to be considered for SDC. Risk of identification is higher for population data than for sample data.
6. **hierarchies/clusters**. Example:
   - Collecting information from all persons in the household.

Types of characteristics in data (2/2)

3. **Sensitive attributes** (e.g. sickness status, costs, late payments, mental disorder, . . .): Information that individuals do not want to be associated with.
4. **Linked** (also sometimes called *Ghost* variables), example: If region is suppressed, the information on municipality must also be suppressed.
5. **weight vector**: If available. Sample weights have to be considered for SDC. Risk of identification is higher for population data than for sample data.
6. **hierarchies/clusters**. Example:
   - Collecting information from all persons in the household.

7. *rest*

How a re-identification is done

... and disclosure scenarios

Types of re-identification

1. **Identity disclosure**. Link of the record with **external data** so that person is identified.

- Example from before: Persons including information about mental disorder. **Record Linkage** of the quasi-identifiers (e.g. age, gender, occupation, municipality) with data from GfK containing names. If the link for a person is successful, the data attacker now knows the names of persons having mental disorder.

Types of re-identification

2. **Attributes Disclosure**.

- Example: A medical study publishes statistics in which **all** people with Austrian nationality between 45 and 50 have dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\longrightarrow$ we learn: Every **individual** Austrian in age group [45-50] living in Winterthur has dementia.

Types of re-identification

2. **Attributes Disclosure**.

- Example: A medical study publishes statistics in which **all** people with Austrian nationality between 45 and 50 have dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\longrightarrow$ we learn: Every **individual** Austrian in age group [45-50] living in Winterthur has dementia.

3. **Inferential Disclosure**. model-based estimation of the value of a sensitive variable: when the quality of prediction is too high.

Re-identification scenarios

1. Nosy neighbour scenario
   - The data recipient has detailed personal information about a specific (or some) person(s).
   - Example: Celebrities in NYC Taxi, tip
2. the archive (matching) scenario
   - Match via key variables with other data sources ("Archives' ') which contain clear names or ID's (*Record Linkage* problem)
   - Re-identify people through successful matches

...

- There are more, but they are less common

Disclosure Risk, general

The most important and complicated part of SDC is not to apply anonymisation methods, but the measurement of the re-identification risk of individuals.

- for register/population data, risk determination is easier.
- non-trivial for survey samples and/or for data with missing values

2 steps:

1. determine the **disclosure scenario** (What are the key variables?) = which overlapping variables are contained in accessible external data sets and can be used for matching (GfK data, BFS data, social media data, ...)
2. **Risk measurement** using SDC methods

Disclosure Risk, general

The most important and complicated part of SDC is not to apply anonymisation methods, but the measurement of the re-identification risk of individuals.

- for register/population data, risk determination is easier.
- non-trivial for survey samples and/or for data with missing values

2 steps:

1. determine the **disclosure scenario** (What are the key variables?) = which overlapping variables are contained in accessible external data sets and can be used for matching (GfK data, BFS data, social media data, ...)
2. **Risk measurement** using SDC methods

Function `sdcMicro::createSdcOjb`

- Function arguments `keyVars`, `numVars`, `weightVar`, ...

sdcMicro::createSdcObj

```
library(sdcMicro)
args(createSdcObj)
```

```
## function (dat, keyVars, numVars = NULL, pramVars = NULL, ghos
##     weightVar = NULL, hhId = NULL, strataVar = NULL, sensible
##     excludeVars = NULL, options = NULL, seed = NULL, randomiz
##     alpha = 1)
## NULL
```

```
?createSdcObj; ?testdata
```

sdcMicro::createSdcObj

```
library(sdcMicro)
args(createSdcObj)
```

```
## function (dat, keyVars, numVars = NULL, pramVars = NULL, ghos
##     weightVar = NULL, hhId = NULL, strataVar = NULL, sensible
##     excludeVars = NULL, options = NULL, seed = NULL, randomiz
##     alpha = 1)
## NULL
```

```
?createSdcObj; ?testdata
```

Define the disclosure scenario:

```
testdata$relat <- as.factor(testdata$relat) # needed afterwards
testdata$roof <- as.factor(testdata$roof) # needed afterwards
sdc <- createSdcObj(testdata,
  keyVars=c('urbrur','relat','sex','age','hhcivil'),
  numVars=c('expend','income','savings'),
  w='sampling_weight',
  pramVars = "roof") # switch to R, explanation S4 class
```

Basic terms Disclosure Risk for populations

Concept of the **Uniqueness**:

- By combining several variables (the QID's), an individual can uniquely can be identified in the data record.
- A key is unique if its frequency is 1 (only one person has the combination of characteristics defined by the key. Example: the key Postcode **8404**, citizienship **Austria**, **male**, **age 45**)

Basic terms Disclosure Risk for populations

Concept of the **Uniqueness**:

- By combining several variables (the QID's), an individual can uniquely can be identified in the data record.
- A key is unique if its frequency is 1 (only one person has the combination of characteristics defined by the key. Example: the key Postcode **8404**, citizienship **Austria**, **male**, **age 45**)

Concept of *k*-**anonymity**:

- Each combination of key variables contains at least *k* observations
- Often we want to ensure 3-anonymity

Basic concepts of risk in data from sample surveys

Example: survey on social contact behaviour in Covid-19 times

- Concept of the **Re-Identification Risk**:
  - Search for rare combinations in the population taking into account the sampling weights of the observations.
  - Difficulty: Frequency of the key is usually not known and must be estimated on the basis of a model.
- A sample in itself already contributes to anonymiztion
  - The data attacker cannot be sure whether a person is in the sample.
  - This is taken into account when estimating the risk.

Risk estimation is generally a difficult mathematical problem, but it is well represented in software.

Risk assessment - overview

- Determine identification risk for each individual in the data set
- Global risk of a data set, e.g. sum of individual risks
- Risk estimation: distinction between **categorical** key variables (such as age, gender, region, ...) and **continuous** key variables (such as costs, income, ...)

Risk assessment - overview

- Determine identification risk for each individual in the data set
- Global risk of a data set, e.g. sum of individual risks
- Risk estimation: distinction between **categorical** key variables (such as age, gender, region, ...) and **continuous** key variables (such as costs, income, ...)

Data of the entire **population** (e.g. data from all persons having diagnosed mental disorded in the Canton of Zurich)

- Concept of Uniqueness, k-anonymity
- l-diversity
- *uniqueness on subsets* (SUDA)

Risk assessment - overview

- Determine identification risk for each individual in the data set
- Global risk of a data set, e.g. sum of individual risks
- Risk estimation: distinction between **categorical** key variables (such as age, gender, region, ...) and **continuous** key variables (such as costs, income, ...)

Data of the entire **population** (e.g. data from all persons having diagnosed mental disorded in the Canton of Zurich)

- Concept of Uniqueness, *k*-anonymity
- *l*-diversity
- *uniqueness on subsets* (SUDA)

Data from complex **surveys**

- individual risk approach
- global risk via log-linear models

Example: k-anonymity and l-diversity

	key variables		f_k	sensitive variable	distinct l-diversity
	gender	age group		stage of dementia	
1	male	30s	3	3	2
2	male	30s	3	0	2
3	male	30s	3	0	2
4	female	20s	3	1	1
5	female	20s	3	1	1
6	female	20s	3	1	1

l-diversity is therefore designed for attribute disclosure

sdcMicro::createSdcObj, k-anonymity

```r
print(sdc, "kAnon")
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 289 (6.310%)
##   - 3-anonymity: 483 (10.546%)
##   - 5-anonymity: 717 (15.655%)
##
## ----------------------------------------------------
```

Special uniques detection algorithm (SUDA)

The so called SUDA scores are more complicated to explain, therefore only the idea:

- An observation is *special unique* with respect to a set of variables Q (e.g. age, sex, place of residence) if it is unique in Q and in a subset of variables of Q (e.g. age, place of residence).
- *Minimal Sample Uniques* (MSUs): unique variable sets with no uniqueness in subsets of these.

Special uniques detection algorithm (SUDA)

- SUDA scores:

1. the smaller the number of variables that span an MSU, the greater the risk of re-identification of the observation
   - example: An observation is already unique in the combination of age and sex $\rightarrow$ risk is higher than if an observation becomes unique only when adding residence.
2. the more MSUs an observation has, the greater the risk of the observation.
   - example: An observation is unique in the combination of age and sex and also in age and place of residence $\rightarrow$ risk is higher than if an observation is unique in age and place of residence only, but not in age and sex.

sdcMicro::createSdcObj, suda scores

```r
sdc <- suda2(sdc)
slot(sdc, "risk")$suda2
```

```
##
## Dis suda scores table:
## - - - - - - - - - - -
##      Interval Number of records
## 1        == 0              4291
## 2  (0.0, 0.1]               281
## 3  (0.1, 0.2]                 8
## 4  (0.2, 0.3]                 0
## 5  (0.3, 0.4]                 0
## 6  (0.4, 0.5]                 0
## 7  (0.5, 0.6]                 0
## 8  (0.6, 0.7]                 0
## 9       > 0.7                 0
```

The individual risk approach for complex surveys

Example, representative sample:

- 5 women living in Winterthur aged 90-100 years (from a total of 500) took part of the questionaire. The design weight would be 100, so without further calibrations the sampling weight would also be 100.
- 5 men living in Winterthur aged 90-100 years answered out of a total of 10. The design/sample weight would therefore be 2.
- in this example, it is easier to identify a men than a woman, although the same number of Winterthur men and women have answered the questionaire.

If one works with surveys including sampling weights, k-anonymity and suda should not be used.

The individual risk approach for sampling

- The fewer observations belong to a key, the higher the risk. More likely to correctly match the observation with external data.
- The smaller a sample weight, the higher the risk.
- Individual risk can be interpreted as **the probability of re-identifying an individual** or as **the probability of a successful match with individuals from external data sources**.

sdcMicro:createSdcObj, individual risk

```
slot(sdc, "risk")$individual %>% head
```

```
##                risk fk   Fk
## [1,] 0.0007686395 14 1400
## [2,] 0.0006246096 17 1700
## [3,] 0.0001723841 59 5900
## [4,] 0.0001639076 62 6200
## [5,] 0.0009990010 11 1100
## [6,] 0.0011098779 10 1000
```

```
riskyCells(sdc, maxDim = 5, threshold = 3) %>% tail
```

```
##      dim1  dim2 dim3    dim4    dim5 threshold unsafe_cells
## 1: urbrur relat  sex     age    <NA>         3          335
## 2: urbrur relat  sex hhcivil    <NA>         3           25
## 3: urbrur relat  age hhcivil    <NA>         3          329
## 4: urbrur   sex  age hhcivil    <NA>         3          251
## 5:  relat   sex  age hhcivil    <NA>         3          301
## 6: urbrur relat  sex     age hhcivil         3          428
```

Disclosure risk for continuous key variables (e.g. income)

- Attacker matches *his* data with published data via overlapping **continuous** variables $\rightarrow$ *record linkage* issue.
- Determining the risk of successfully matched individuals.

Disclosure risk for continuous key variables (e.g. income)

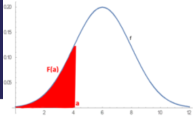
- Attacker matches *his* data with published data via overlapping **continuous** variables $\rightarrow$ *record linkage* issue.
- Determining the risk of successfully matched individuals.

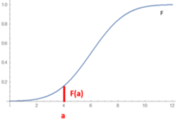
```
# only makes sense after anonymization
slot(sdc, "numrisk")
```

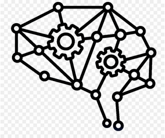
Methods for anonymisation of data

Different groups of methods:

- Methods that generalize or suppress values. Examples are recoding or local suppression
- Methods which perturb data. Examples are Adding Noise, Post-Randomization Method (PRAM), Microaggregation and Shuffling.
- Methods for generating synthetic data





Recoding

Recoding of categorical key variables:

- achieve anonymity by merging/generalising categories
  - Example: Combining / generalising several postal codes (8400, 8401, 8402, 8403, 8404 to 840x)

Recoding continuous variables

- means to discretise the variable
- (limping) example: exact age of a person to age categories

Recoding

```
?groupAndRename
?globalRecode
```

```
sdc <- globalRecode(sdc,
                    column="age",
                    breaks=c(1,9,19,29,39,49,59,69,100))
print(sdc, "kAnon")
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 51 (1.114%) | in original data: 289 (6.310%)
##   - 3-anonymity: 98 (2.140%) | in original data: 483 (10.546%
##   - 5-anonymity: 164 (3.581%) | in original data: 717 (15.655
##
## ----------------------------------------------------------------
```

```
# print(sdc, "risk")
```

Recoding

```
sdc <- groupAndRename(sdc,
                      var="relat",
                      before=1:9,
                      after=c(1:6,"7+","7+","7+"))
# print(sdc, "kAnon")
print(sdc, "risk")
```

```
## Risk measures:
##
## Number of observations with higher risk than the main pa
##   in modified data: 0
##   in original data: 0
## Expected number of re-identifications:
##   in modified data: 3.47 (0.08 %)
##   in original data: 18.46 (0.40 %)
```

Local suppression

Problem: with recoding, the risk has been significantly reduced, but some people still have an increased risk. If further information was recoded, the quality of data analysis would suffer too much.

Local suppression

- **Aim**: to suppress values as little as possible and to guarantee e.g. k-anonymity (find an optimal suppression pattern)
- Typically used **after** a recoding to minimise the residual risk.
- Heuristic **optimisation methods** to find specific patterns in categorical key variables. Replace this pattern with missing values.
- Further complexity: Frequencies of keys with missing values.
- Weighting of variables according to their importance

sdcMicro::kAnon

```
sdc <- kAnon(sdc, k = 3, importance = c(3,4,1,2,5))
print(sdc, "kAnon")
```

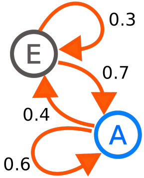
```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 0 (0.000%) | in original data: 289 (6.310%)
##   - 3-anonymity: 0 (0.000%) | in original data: 483 (10.546%)
##   - 5-anonymity: 46 (1.004%) | in original data: 717 (15.655%)
##
## -------------------------------------------------------------------
```

```
print(sdc, "risk")
```

```
## Risk measures:
##
## Number of observations with higher risk than the main part of the da
##   in modified data: 0
##   in original data: 0
## Expected number of re-identifications:
##   in modified data: 0.88 (0.02 %)
##   in original data: 18.46 (0.40 %)
```

Post Randomization (PRAM)

- Swap values between categories of a variable with given transition probabilities.
  - Example: with a probability of 0.1, the place of residence Oberwinterthur is swapped with the place of residence Winterthur-Hegi.
  - In practice mostly within strata (e.g. swapping a person's postcode only within a canton).
- Attacker can never be sure whether a value is true or has been swapped.
- Popularly used in practice: swap geographical information with PRAM



sdcMicro:createSdcObj, PRAM

```
sdc <- pram(sdc)
print(sdc, "pram")
```

```
## Post-Randomization (PRAM):

## Variable:roof

## --> final Transition-Matrix:

##                2          4            5            6
## 2 0.960116368 0.03881884 0.0002501457 0.0003728117 0.000
## 4 0.008547074 0.98492463 0.0019230057 0.0029242904 0.001
## 5 0.010716767 0.37417642 0.6114796782 0.0007641125 0.002
## 6 0.008925550 0.31797357 0.0004270041 0.6702748441 0.002
## 9 0.022478559 0.38841600 0.0033998350 0.0050979387 0.580

##
## Changed observations:

##    variable nrChanges percChanges
```

Anonymization of continous key variables

Continuous key variables usually not present in data in the research area of psychlogy, thus only some names of methods

- **Microaggregation**: find similar observations (clustering problem) and replace the values with an aggregate (e.g. arithm. mean)
- **Adding Noise**: e.g. add random noise to year of born or income ...
- **Shuffling**: more complex method uses a statistical (regression) model, but with some flaws.
- ...

sdcMirco::microaggregation, addNoise, shuffle

```
sdc <- addNoise(sdc, method = "correlated2")
print(sdc, "numrisk")
```

```
## Numerical key variables: expend, income, savings
##
## Disclosure risk (~100.00% in original data):
##   modified data: [0.00%; 5.55%]
##
## Current Information Loss in modified data (0.00% in orig
##   IL1: 475885.91
##   Difference of Eigenvalues: 0.400%
## -----------------------------------------------------------
```

sdcMirco::microaggregation, addNoise, shuffle

```
sdc <- undolast(sdc)
sdc <- addNoise(sdc, method = "additive", noise = 10)
sdc <- dRiskRMD(sdc)
slot(sdc, "risk")$numericRMD$wrisk2
```

```
## [1] 0.1188653
```

```
print(sdc, "numrisk")
```

```
## Numerical key variables: expend, income, savings
##
## Disclosure risk (~100.00% in original data):
##   modified data: [0.00%; 6.29%]
##
## Current Information Loss in modified data (0.00% in orig
##   IL1: 521998.83
##   Difference of Eigenvalues: 0.430%
## ----------------------------------------------------------
```

Utility

- After data has been anonymised, it is important to assess the **information loss** and the **data quality**.
- Comparing results from original and anonymised data (tables, regression models, distributions, . . .)
- Comparison of indicators
- Propensity score matching methods
- Etc.

If the loss of data is high, anonymisation should be considered.

Trade-Off and iterative approach (Anonymisation $\leftrightarrow$ Utility)

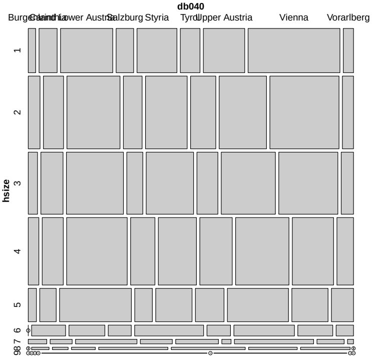
Utility

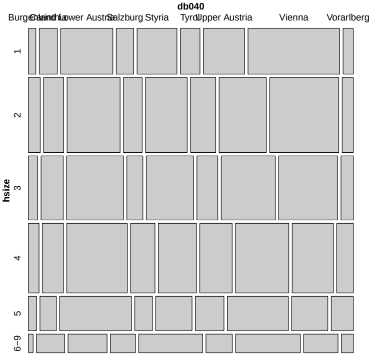
```r
print(sdc, "ls")
```

```
## Local suppression:

##   KeyVar | Suppressions (#) | Suppressions (%)
##   urbrur |                2 |            0.044
##    relat |               36 |            0.786
##      sex |                0 |            0.000
##      age |                2 |            0.044
##  hhcivil |               56 |            1.223

## ----------------------------------------------------
```

More in R script . . .

Utility Comparison of tables, visual





Mosaic plot of gender (rb090) × citizenship (pb220a) × household size (hsize) with the original sampling frequencies (left diagram) and the sampling frequencies from the anonymised data (right diagram).

Software

sdcMicro (Templ et al., Journal of Statistical Software, 2016)

- state-of-the-art software
- can handle more complex data
- with click-App (for the browser)
- is programmed very efficiently (C++ code, parallel computing)

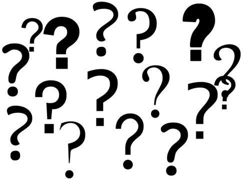
simPop (Templ et al., Journal of Statistical Software, 2017)

- for the creation of synthetic data sets
- unlike other software, can also handle more complex data structures

sdcTable and **cellKey** (Author: B. Meindl)

- For the confidentiality of tables (aggregated information)

Difficulties in practice



- Unfortunately there is no general solution and no standardised procedure
- Anonymisation varies from case to case. Strongly data- and case-dependent
- Years of experience necessary

News

Fellowship DIZH *Anonymisation and estimation of the re-identification risk of personal data*, Competence Centre Data Anonymisation.

- Start Fellowship: Sept. 2020.
- Anonymisation lab will be founded in 2021/22.