Matthias Templ

Statistik Austria
Vienna Univ. of Techn.

Workshop on
Anonymization tools and
their practical relevance
2015

The sdcMicro and sdcMicroGUI packages for statistical disclosure control
Introduction

- anonymisation of data sets often required due to given laws on privacy
- it’s about balancing **user needs** (ideal case: original data) and **privacy** (ideal case: no data)
- the goal is to provide datasets for release that don’t allow users to link information to specific individuals/enterprises
Linking Information - a different view ;-) 

By linking external information → we possibly learn something about the wheelchair user. Linking may not be successful for the others.
What is Statistical Disclosure Control?

**Disclosure:**
Someone learns something about someone that was not previously known using released data

Statistical Disclosure Control **methods** are in cope with

- ... perturbation of real complex data (in `sdcMicro`)
- ... simulation of synthetic data (in `simPop`)
- ... measuring disclosure risk (in `sdcMicro`)
- ... comparing original and modified data (information loss/data utility) (in `sdcMicro/R`)
- ... protecting multidimensional linked tabular data (in `sdcTable`)

Templ (sdcMicro)
What are the real problems?

- **Huge data** sets and the need of efficient algorithms and implementations
- **Complex structures**, data sampled with complex designs
- **Missing values** and structural zeros
- Compositional nature of components with high amount of zeros
Concepts

- Re-identification may occur due to
  - direct identifying variables
  - indirect identifying variables
- direct identifiers have to be removed from the sample
- indirect identifying variables are usually
  - publicly available information or
  - available in public databases

**categorical key variables:**
categorical indirect identifying variables $\rightarrow$ **cross-classification** of them determines the key's
Quantifying risk is based on the distribution of the keys

- in the sample (example: is the combination of key(i) State = AUT, Ethnicity = Korean, Age = 50, Gender = F, Occupation = University Lecturer unique in the sample?)

- in the sample and in the population (example cont’d: how many people exists in the population with key(i) ? → if, e.g. 3 and linking is possible, the intruder have probability 1/3 that it is the correct link.)
Various methods exist/are implemented. For **categorical** variables:

- **k-anonymity concept** (frequency of each key $> k$ in the sample)
- **SUDA** (also sample-based, but consider subsets of keys)
- **Individual risk** based on superpopulation models (distribution on frequencies in the population modelled)

For **continuous** variables:

- **distance-based methods**
- probabilistic methods
Aim: \textit{k-anonymity} and low disclosure risk (individual risk, global risk, suda2)

- deterministic methods
  - top- and bottom coding
  - recoding
  - (optimal or risk-based) local suppression

- probabilistic methods
  - (rank) swapping
  - post-randomization (pram)
Aim: perurbe data so that linking is not successful

- deterministic protection methods
  - top- and bottom coding
  - microaggregation (most similar observations are aggregated)

- perturbative protection methods based on randomness
  - adding correlated noise (take the covariance structure into account)
  - (sampling)
  - (rank)swapping (swapp values within an pre-defined range)
  - shuffling (model-based)
Simplified Workflow

1. delete direct identifier
2. define key variables
   - simulate fully synthetic data
   - simulate partly synthetic data
3. measure risk
   - for continuous variables
     - adding noise
     - microaggregation
     - Shuffling
4. Recoding
   - for categorical variables
   - local suppression
5. measure information loss
   - if utility is acceptable high
   - if risk is small and utility is high
6. release data
The Software: sdcMicro

- The S4-class oriented R-package sdcMicro contains all methods, and a point-and-click user-interface (package sdcMicroGUI) allows to apply the methods also without knowledge in R.
- Once a sdcMicroObj is defined, anonymisation is straightforward.
- Efficiently programmed. Even data with millions of observations can be processed.
The Software: sdcMicro

- The S4-class oriented R-package *sdcMicro* contains all methods, and a point-and-click user-interface (package *sdcMicroGUI*) allows to apply the methods also without knowledge in R.
- Once a *sdcMicroObj* is defined, anonymisation is straightforward.
- Efficiently programmed. Even data with millions of observations can be processed.
The Software: sdcMicro

- The S4-class oriented R-package sdcMicro contains all methods, and a **point-and-click** user-interface (package sdcMicroGUI) allows to apply the methods also without knowledge in R.
- Once a sdcMicroObj is defined, anonymisation is straightforward.
- Efficiently programmed. Even data with millions of observations can be processed.
sdcMicro: Brief Example

Import, for example, Bangladeshi income data (48969 x 41) into R:

```r
bgd05 <- read.dta("BGD_2005_I2D2.dta")
```

Define a `sdcMicro` object once:

```r
sdc <- createSdcObj(dat=bgd05, 
  keyVars=c('gender','age','marital',
            'empstat','reg01'),
  weightVar = 'wgt',
  hhId = 'idh',
  numVar=c("wage","pci","pcc"))
```
Import, for example, Bangladeshi income data (48969 x 41) into R:

```r
bgd05 <- read.dta("BGD_2005_I2D2.dta")
```

Define a sdcMicro object once:

```r
sdctemp <- createSdcObj(dat=bgd05,
                         keyVars=c('gender','age','marital',
                                   'empstat','reg01'),
                         weightVar = 'wgt',
                         hhId = 'idh',
                         numVar=c("wage","pci","pcc"))
```
The S4-class sdcMicroObj

```r
slotNames(sdc)

## [[1]] "origData"  "keyVars"  "pramVars"
## [[4]] "numVars"   "weightVar"  "hhId"
## [[7]] "strataVar" "sensibleVar" "manipKeyVars"
## [[10]] "manipPramVars" "manipNumVars" "manipStrataVar"
## [[13]] "originalRisk" "risk"           "utility"
## [[16]] "pram"     "localSuppression" "options"
## [[19]] "additionalResults" "set"      "prev"
## [[22]] "deletedVars"
```
### Number of observations violating

- 2-anonymity: 539
- 3-anonymity: 1031

### Percentage of observations violating

- 2-anonymity: 1.1%
- 3-anonymity: 2.11%
print(sdc, "risk")

##
## --------------------------
## 0 obs. with higher risk than the main part
## Expected no. of re-identifications:
##  2.88 [ 0.01 %]
## --------------------------
## Hierarchical risk
## --------------------------
## Expected no. of re-identifications:
## 15.82 [ 0.03 %]
in general:

\[ \text{sdcObj} \leftarrow \text{method(sdcObj)} \]

and for summaries and plots:

\[ \text{print(sdcObj)} \]
\[ \text{plot(sdcObj)} \]
sdcMicro: Recoding

summary(sdc@origData$age)

```
# #  Min. 1st Qu. Median   Mean  3rd Qu.   Max.
# # 0.00 10.00 21.00  25.54  38.00  98.00
```

Recoding with `globalRecode()` or `groupVars()`

```r
dc <- globalRecode(sdc,
                   column="age",
                   breaks=c(-1,9,19,29,39,49,59,69,130),
                   labels=paste("age",1:8,sep=""))
```

summary(extractManipData(sdc)$age)

```
# #  age1  age2  age3  age4  age5  age6  age7  age8  
# # 11787 11190  7677  6690  5185  3169  1934  1337
```
Recoding with `globalRecode()` or `groupVars()`

```r
sdc <- globalRecode(sdc, 
  column="age", 
  breaks=c(-1,9,19,29,39,49,59,69,130), 
  labels=paste("age", 1:8, sep=""))

summary(extractManipData(sdc)$age)
```

```r
## age1  age2  age3  age4  age5  age6  age7  age8
## 11787 11190  7677  6690  5185  3169  1934  1337
```
ensuring 3-anonymity: (finds minimal number of values to suppress)

```r
sdc <- localSuppression(sdc, k=3)
print(sdc, "risk")
```

##
## --------------------------
## 0 (orig: 0 ) obs. with higher risk than the main part
## Expected no. of re-identifications:
## 0.11 [ 0 %] (orig: 2.88 [ 0.01 %])
## --------------------------
## --------------------------
## Hierarchical risk
## Expected no. of re-identifications:
## 0.61 [ 0 %] (orig: 15.82 [ 0.03 %])
sdc <- microaggregation(sdc, strata_variables="strata")
print(sdc, "numrisk")

## Disclosure Risk is between:
## [0% ; 94.98%] (current)
##
## (orig: ~100%)
## - Information Loss:
##   IL1: 0.02
## - Difference Eigenvalues: 7.36 %
##
## (orig: Information Loss: 0)
```r
dc <- undolast(dcc)  # undo previous anon

sdc <- shuffle(sdc, form = wage + pcc + pci ~ age + marital + empstat + gender)

print(sdc, "numrisk")
```

## Disclosure Risk is between:
## [0% ; 16.82%] (current)
##
## (orig: ~100%)

## - Information Loss:
##   IL1: 0.64
## - Difference Eigenvalues: 71.66 %
##
## (orig: Information Loss: 0)
sdcMicro: A lot of print methods...

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

```r
## Reported is the
## number | mean size and | size of smallest category
## ------------
## gender ... 2 | 24484 | 24314
## (orig: 2 | 24484 | 24314)
## ------------
## age ...... 9 | 6114 | 1332
## (orig: 99 | 495 | 1)
## ------------
## marital .. 5 | 12174 | 319
## (orig: 5 | 12174 | 319)
## ------------
## empstat .. 5 | 3318 | 773
## (orig: 5 | 3318 | 773)
## ------------
## reg01 .... 7 | 6995 | 0
## (orig: 7 | 6995 | 0)
```
sdcMicro: A lot of print and plot methods...

```r
print(sdc, "ls")
print(sdc, type="pram")
...
```

(suppressing output)

and reports

```r
args(report)
```

```r
## function (obj, outdir = getwd(), filename = "SDC-Report",
##     title = "SDC-Report", internal = FALSE)
## NULL
```

```r
report(sdc)
report(sdc, internal=TRUE)
...
```
sdcMicroGUI - easy to use graphical interface

(a) The *data* menu entry.

(b) On-the-fly preview of *.csv* files.

importing csv files.
sdcMicroGUI - Tab Identifiers

Loaded data set:
testdata [n=4580]

Selected key variables

Categorical
- urbrur [#:2]
- roof [#:5]
- walls [#:3]
- water [#:8]

Numerical
- expend [Min: 3377, Med: 50462299.5, Max: 99962044]
- income [Min: 2897.5, Med: 50750000, Max: 1e+08]
- savings [Min: 2974.5, Med: 4982921, Max: 9997808]

Selected auxiliary variables

Weight
- sampling_weight [Min:100, Med:100, Max:100]

Household ID
- ori_hid [Mean size:4.6]

Strata
- not selected
sdcMicroGUI - Tab Categorical

- Frequency calculations
  - Number of observations violating
    - 2-anonymity: 1 (orig: 1)
    - 3-anonymity: 5 (orig: 5)
  - Percentage of observations violating
    - 2-anonymity: 0.02% (orig: 0.02%)
    - 3-anonymity: 0.11% (orig: 0.11%)

- Risk for categorical key variables
  - 0 (orig: 0) obs. with higher risk than the main part
  - Expected no. of re-identifications: 0.56 [0.01%] (orig: 0.56 [0.01%])
  - Hierarchical risk
    - Expected no. of re-identifications: 2.66 [0.06%] (orig: 2.66 [0.06%])

- View Observations with high risk
- I-Diversity

- Protection
  - Recode
  - pram
  - Local suppression
    - optimal (k-Anonymity)
    - threshold (indiv.Risk)

- Information Loss
  - Recodings
    - Reported is the number | mean size and | size of smallest category
    - urbrur .. 2 | 2290 | 646
      (orig: 2 | 2290 | 646)
    - roof .... 5 | 916 | 16
      (orig: 5 | 916 | 16)
    - walls ... 3 | 1527 | 50
      (orig: 3 | 1527 | 50)
    - water .. 8 | 572 | 26
      (orig: 8 | 572 | 26)
  - Suppressions
    - urbrur .. 0 [0%]
    - roof .... 0 [0%]
    - walls ... 0 [0%]
    - water ... 0 [0%]
sdcMicroGUI - Example: Recode
sdcMicroGUI - Example: Recode
sdcMicroGUI - reproducibility

```
Row.names items
1 activedataset <- testdata
2 sdcObject <- createSdcObj(activedataset,keyVars=c('urbrur', 'roof', 'walls', 'water'), numVars=c('expend', 'income', 'save'))
3 sdcObject <- varToFactor(sdcObject, var=c('roof'))
4 sdcObject <- varToFactor(sdcObject, var=c('roof'))
5 sdcObject <- varToFactor(sdcObject, var=c('roof'))
6 sdcObject <- varToFactor(sdcObject, var=c('urbrur'))
7 sdcObject <- varToFactor(sdcObject, var=c('walls'))
8 sdcObject <- varToFactor(sdcObject, var=c('water'))
9 sdcObject <- groupVars(sdcObject, var=c('water'), before=c('6', '7', '9'), after=c('6;7;9'))
10 sdcObject <- groupVars(sdcObject, var=c('roof'), before=c('5', '6'), after=c('5;6'))
11 sdcObject <- groupVars(sdcObject, var=c('roof'), before=c('5;6', '9'), after=c('5;6;9'))
12 sdcObject <- localSuppression(sdcObject, k=c(3), importance=c(1, 2, 2, 4))
```
log(pcc) ~ age + literacy + edulevel2 + computer + urb + gender + reg01, data=bgd05, weights=wgt

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | 6.7764003  | 0.0111136 | 609.741 < 2e-16 *** |
| age | 0.0028465  | 0.0001197 | 23.776 < 2e-16 *** |
| literacyYes | 0.1144539  | 0.0253004 | 4.524 6.09e-06 *** |
| edulevel2Primary | 0.0575543  | 0.0255432 | 2.253 0.0243 * |
| edulevel2Secondary | 0.2422043  | 0.0254577 | 9.514 < 2e-16 *** |
| edulevel2Post-secondary | 0.5349532  | 0.0268888 | 19.895 < 2e-16 *** |
| computerYes | 0.7795342  | 0.0176839 | 44.082 < 2e-16 *** |
| urbRural | -0.2251389  | 0.0053004 | -42.476 < 2e-16 *** |
| genderFemale | 0.0234892  | 0.0043381 | 5.415 6.17e-08 *** |
| reg01Chittagong | 0.1709304  | 0.0098727 | 17.314 < 2e-16 *** |
| reg01Dhaka | 0.1938753  | 0.0094055 | 20.613 < 2e-16 *** |
| reg01Khulna | -0.0919458  | 0.0105521 | -8.714 < 2e-16 *** |
| reg01Rajshahi | -0.0954591  | 0.0095737 | -9.971 < 2e-16 *** |
| reg01Sylhet | 0.2002829  | 0.0121471 | 16.488 < 2e-16 *** |

Multiple R-squared: 0.3338, Adjusted R-squared: 0.3336
### Model Summary

The model estimated is:

\[ \log(pcc) \sim \text{age} + \text{literacy} + \text{edulevel2} + \text{computer} + \text{urb} + \text{gender} + \text{reg01}, \text{ data=bgd05sdc, weights=wgt} \]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | 6.7756365 | 0.0111401 | 608.220 | < 2e-16 *** |
| age | 0.0028438 | 0.0001202 | 23.667 | < 2e-16 *** |
| literacyYes | 0.1146391 | 0.0253354 | 4.525 | 6.06e-06 *** |
| edulevel2Primary | 0.0569082 | 0.0255783 | 2.225 | 0.0261 * |
| edulevel2Secondary | 0.2420716 | 0.0254928 | 9.496 | < 2e-16 *** |
| edulevel2Post-secondary | 0.5338675 | 0.0269291 | 19.825 | < 2e-16 *** |
| computerYes | 0.7800722 | 0.0176796 | 44.123 | < 2e-16 *** |
| urbRural | -0.2251009 | 0.0053039 | -42.441 | < 2e-16 *** |
| genderFemale | 0.0240010 | 0.0043447 | 5.524 | 3.33e-08 *** |
| reg01Chittagong | 0.1713775 | 0.0099001 | 17.311 | < 2e-16 *** |
| reg01Dhaka | 0.1941658 | 0.0094329 | 20.584 | < 2e-16 *** |
| reg01Khulna | -0.0911513 | 0.0105838 | -8.612 | < 2e-16 *** |
| reg01Rajshahi | -0.0951491 | 0.0096019 | -9.909 | < 2e-16 *** |
| reg01Sylhet | 0.2010148 | 0.0122058 | 16.469 | < 2e-16 *** |

---

Multiple R-squared: 0.334, Adjusted R-squared: 0.3338
Application to Countries - Original Data

The diagram shows the risk measures for various countries. The measures include:

- 3-anon
- ind-risk
- suda2

The countries include:

- BGD05
- IDN10
- IND99
- MEX08
- THA09
- VNM06
- ZAF07

The y-axis represents the risk level, while the x-axis displays the countries.
Computational Costs

The graph shows the relationship between the number of observations and the time in seconds for frequency estimation and risk measurement. The trend line indicates a linear increase in time as the number of observations grows.
Computational Costs

The provided image contains a line graph titled "localSuppression()". The x-axis represents the size, ranging from 0 to 1000000, while the y-axis represents time, ranging from 0 to 100. The graph shows a pattern where the time decreases sharply as the size increases, reaches a minimum, and then increases again as the size continues to increase.

The graph is part of a larger document, indicating it was extracted from a page of a report or presentation.
## Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Software</th>
<th>$\mu$-Argus 4.2</th>
<th>sdcMicro 1.0.0</th>
<th>sdcMicro 4.3.0</th>
<th>sdcMicroGUI 1.1.0</th>
<th>IHSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency counts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>individual risk</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>individual risk on households</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>l-diversity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>suda2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>global risk</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>global risk with log-lin mod.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>recoding</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>local suppression</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>swapping</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>pram</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>adding correlated noise</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>microaggregation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>shuffling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>utility measures</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GUI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CLI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>missing values</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>cluster designs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>large data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>reporting</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>platform independent</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>free and open-source</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

List of methods supported by different statistical disclosure control software. Ticks are in brackets when only limited support is provided to a method. A comparison to version 1.0.0 of sdcMicro (released in May 29, 2008) is given to show the progress of the new complete reimplementation of the package.
Summary

- **sdcMicro** is a well-defined efficient (C++), platform-independent, free and open-source, object-oriented S4-class R package
- **sdcMicroGUI** is a point-and-click user-interface (no R knowledge needed) that is on top of sdcMicro
- reproducible results with CLI and point-and-click GUI

---

- the software was funded by data-analysis OG, the International Household Survey Network, OECD, Worldbank, Google, Statistics Austria and Vienna University of Technology
- sdcMicro and sdcMicroGUI are used on large-scale
- sdcMicro in Journal of Statistical Software (accepted 18.10.2014)
- Springer book in 2015
Summary

- **sdcMicro** is a well-defined efficient (C++), platform-independent, free and open-source, object-oriented S4-class R package
- **sdcMicroGUI** is a point-and-click user-interface (no R knowledge needed) that is on top of sdcMicro
- reproducible results with CLI and point-and-click GUI
- the software was funded by data-analysis OG, the International Household Survey Network, OECD, Worldbank, Google, Statistics Austria and Vienna University of Technology
- sdcMicro and sdcMicroGUI are used on large-scale
- sdcMicro in Journal of Statistical Software (accepted 18.10.2014)
- Springer book in 2015
Summary

- **sdcMicro** is a well-defined efficient (C++), platform-independent, free and open-source, object-oriented S4-class R package

- **sdcMicroGUI** is a point-and-click user-interface (no R knowledge needed) that is on top of sdcMicro

- reproducible results with CLI and point-and-click GUI

- the software was funded by data-analysis OG, the International Household Survey Network, OECD, Worldbank, Google, Statistics Austria and Vienna University of Technology

- sdcMicro and sdcMicroGUI are used on large-scale

- sdcMicro in Journal of Statistical Software (accepted 18.10.2014)

- Springer book in 2015
Summary

- **sdcMicro** is a well-defined efficient (C++), platform-independent, free and open-source, object-oriented S4-class R package
- **sdcMicroGUI** is a point-and-click user-interface (no R knowledge needed) that is on top of sdcMicro
- reproducible results with CLI and point-and-click GUI

- the software was funded by data-analysis OG, the International Household Survey Network, OECD, Worldbank, Google, Statistics Austria and Vienna University of Technology
- sdcMicro and sdcMicroGUI are used on large-scale
- sdcMicro in Journal of Statistical Software (accepted 18.10.2014)
- Springer book in 2015
No one is linkable anymore ...