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 $\rightarrow$ we possible 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)

Challenges

STATISTIK AUSTRIA
Die Informationsmanager

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

- 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 $\longrightarrow$ **cross-classification** of them determines the **key**'s

Risk Assessment for Categorical Key Variables

- 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) ? $\rightarrow$ if, e.g. 3 and linking is possible, the intruder have probability 1/3 that it is the correct link.)

Methods for Risk Estimation

Various methods exists/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)
- Global risk estimated by log-linear models (Clogg and Eliason (1987), Skinner and Holmes (1998), Franconi and Polettini (2004), Shlomo and Skinner (2008), Skinner and Vallet (2010))

For **continuous** variables:

- distance-based methods
- probabilistic methods

Concepts to Anonymize Categorical Variables

Aim: *k*-anonymity **and** low disclosure risk (individual risk, global risk, suda2)

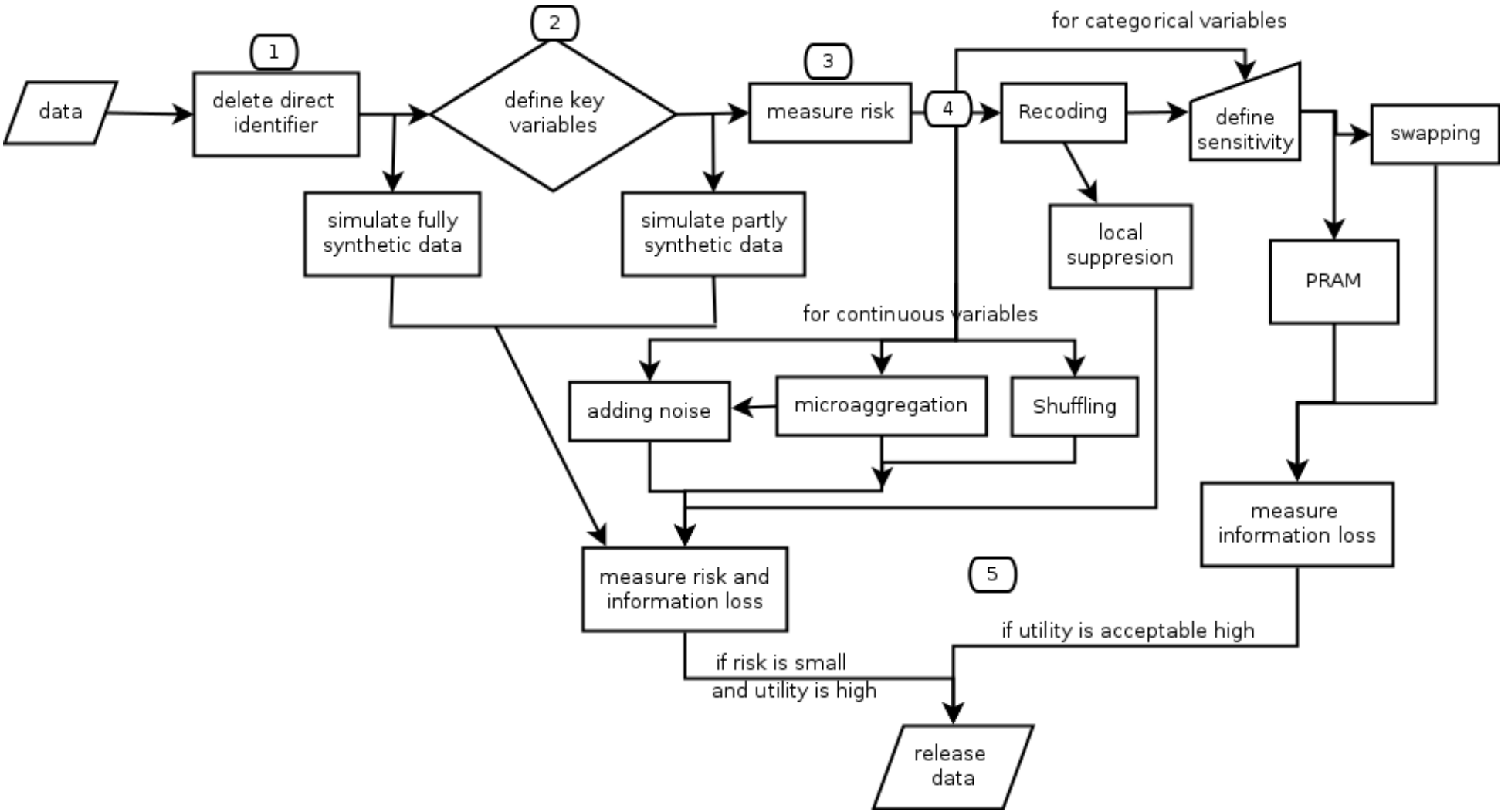
- deterministic methods
  - top- and bottom coding
  - recoding
  - (optimal or risk-based) local suppression
- probabilistic methods
  - (rank) swapping
  - post-randomization (pram)

Concepts to Protect **Continuous** Variables

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



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, Bangaldesh income data (48969 x 41) into R:

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

Define a sdcMicro object once:

```
sdc <-  createSdcObj(dat=bgd05,
                     keyVars=c('gender','age','marital',
                               'empstat',"reg01"),
                     weightVar = 'wgt',
                     hhId = 'idh',
                     numVar=c("wage","pci","pcc"))
```

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

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

Define a sdcMicro object once:

```
sdc <-  createSdcObj(dat=bgd05,
                     keyVars=c('gender','age','marital',
                               'empstat',"reg01"),
                     weightVar = 'wgt',
                     hhId = 'idh',
                     numVar=c("wage","pci","pcc"))
```

The S4-class sdcMicroObj

```
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"
```

sdcMicro: Brief Example

```
print(sdc)

## 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:

```
sdcObj <-  method(sdcObj)
```

and for summaries and plots:

```
print(sdcObj)
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()`

```
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)

##  age1  age2  age3  age4  age5  age6  age7  age8
## 11787 11190  7677  6690  5185  3169  1934  1337
```

```
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()`

```
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)

##  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)

```
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)
```

sdcMicro: Continuous Keys

```
sdc <- undolast(sdc) # 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...

```
print(sdc, "recode")

## 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...

STATISTIK AUSTRIA
Die Informationsmanager

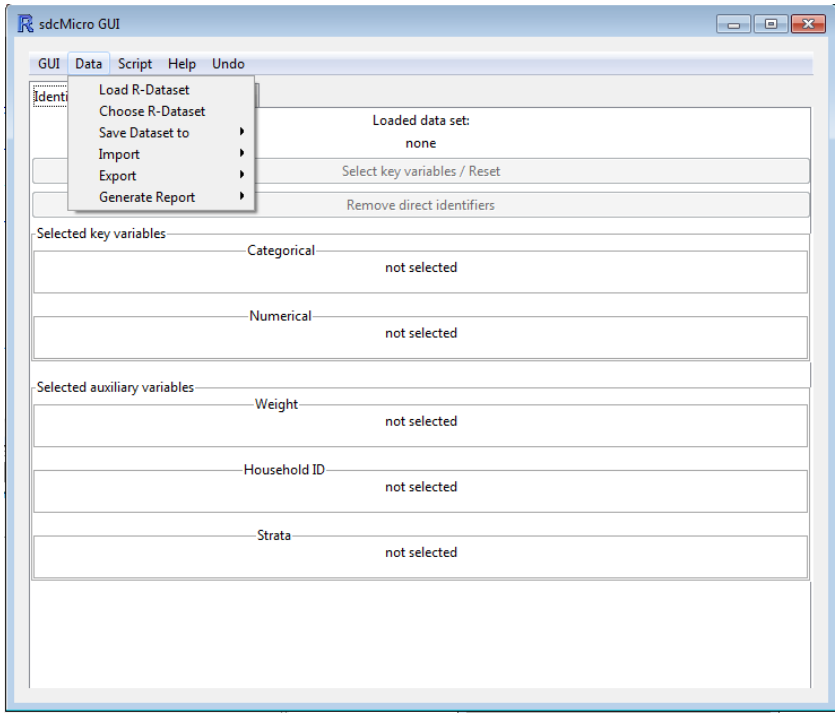
```
print(sdc, "ls")
print(sdc,type="pram")
...
```

(suppressing output)
and reports

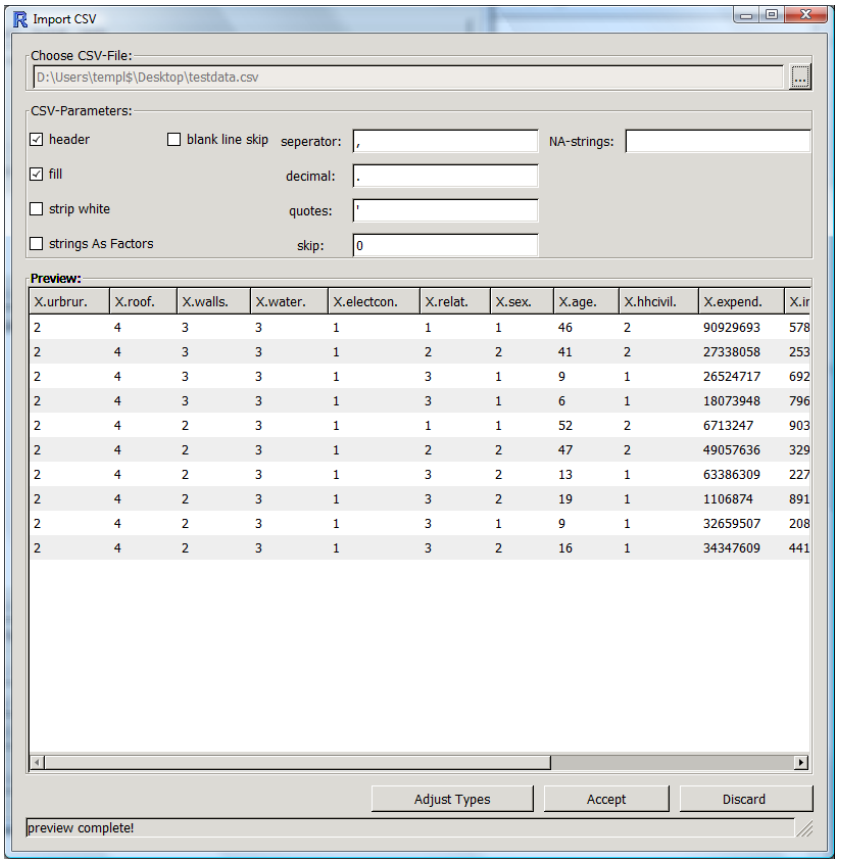
```
args(report)

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

```
report(sdc)
report(sdc, internal=TRUE)
...
```

sdcMicroGUI - easy to use graphical interface



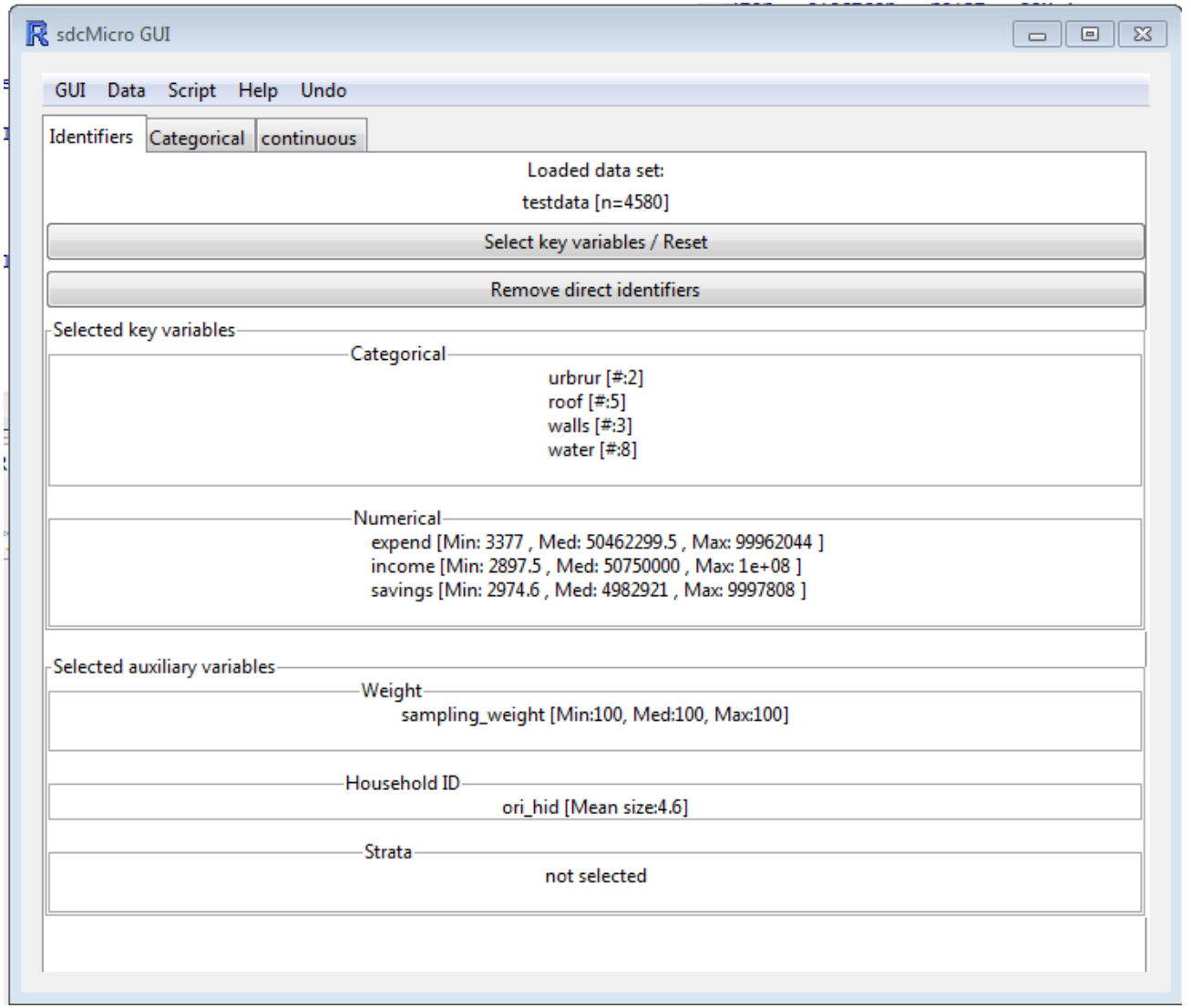
(a) The *data* menu entry.



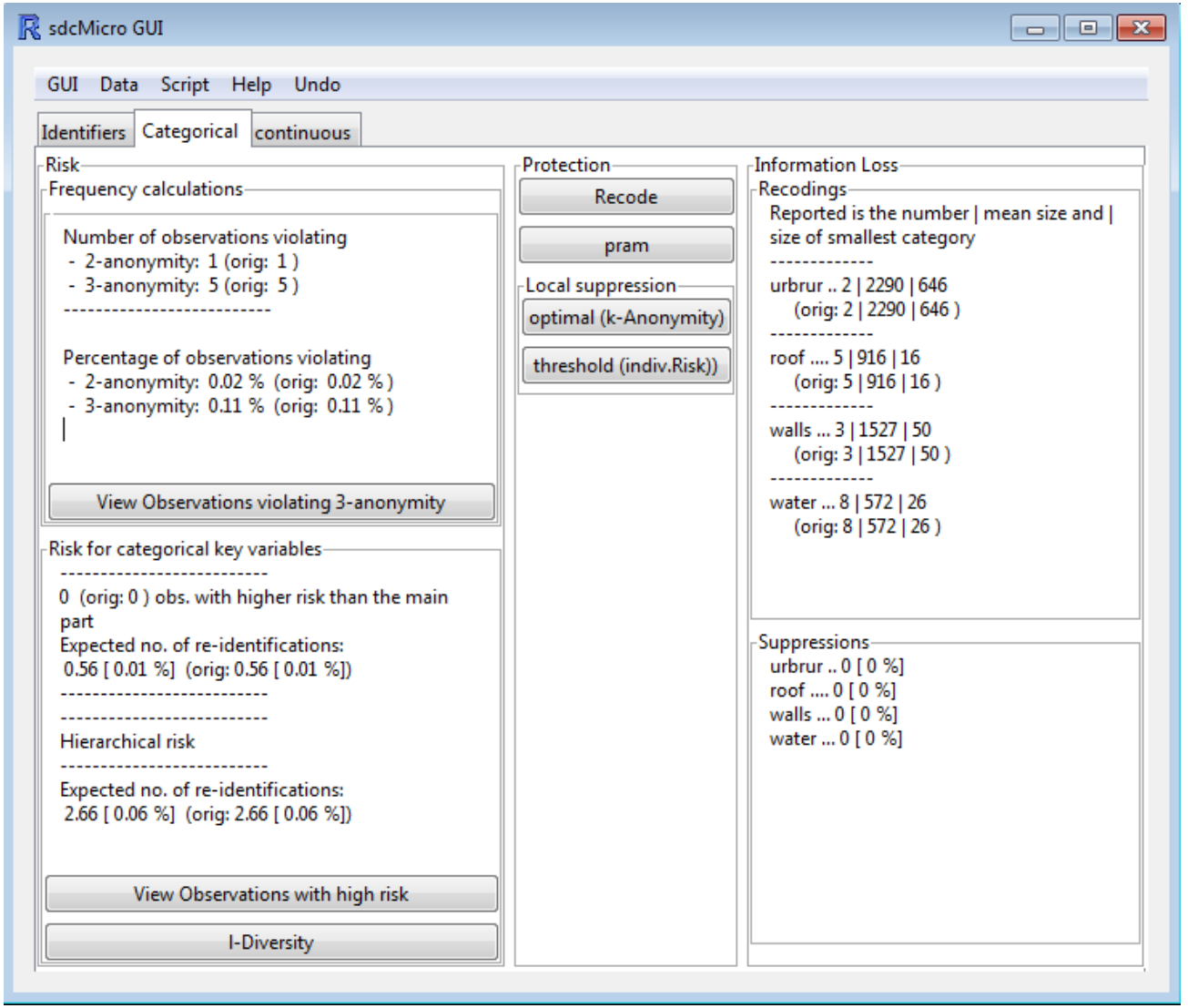
(b) On-the-fly preview of `.csv` files.

importing csv files.

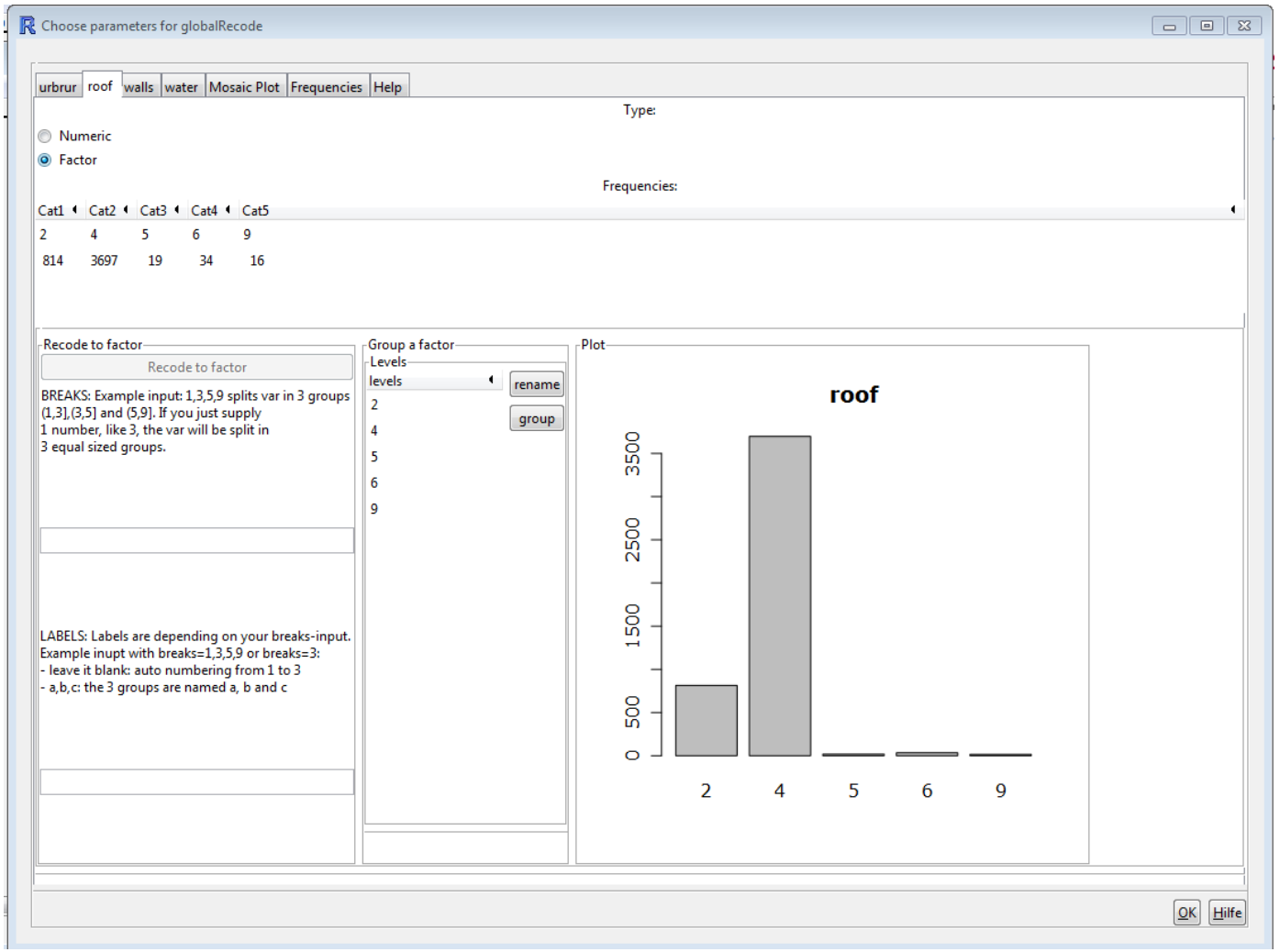
sdcMicroGUI - Tab Identifiers



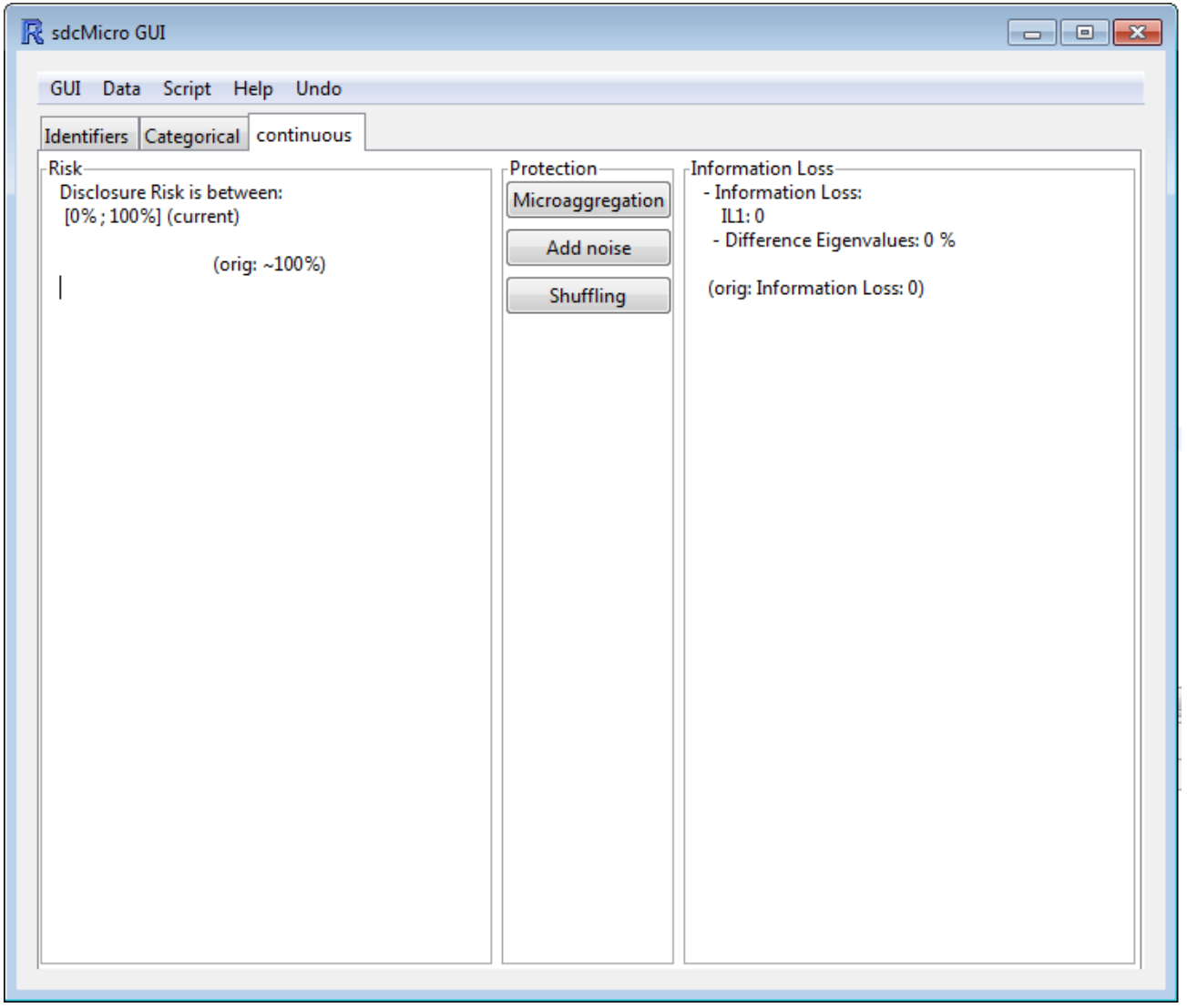
sdcMicroGUI - Tab Categorical



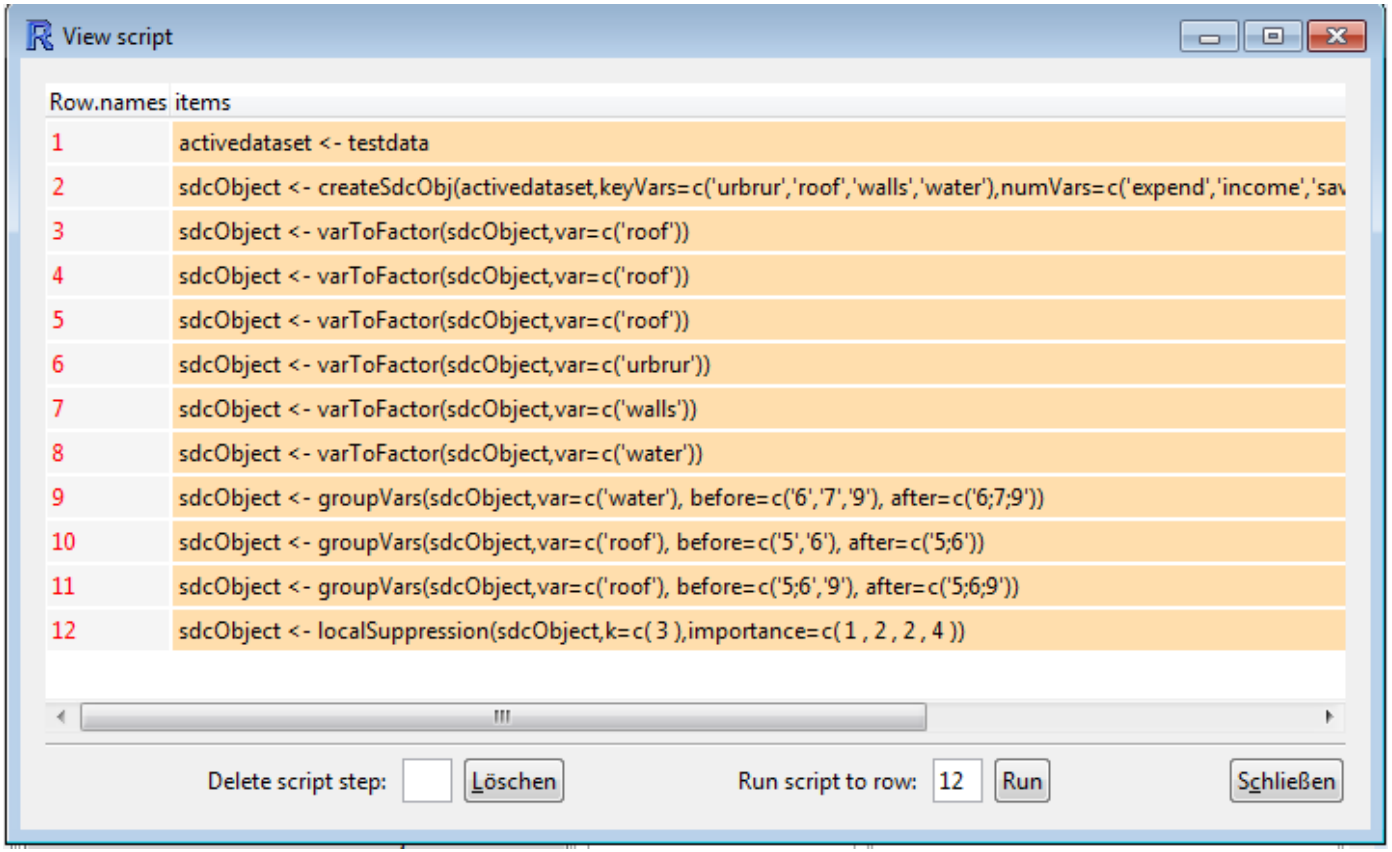
sdcMicroGUI - Example: Recode



sdcMicroGUI - Example: Recode



sdcMicroGUI - reproducibilty



Utility

```
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	***

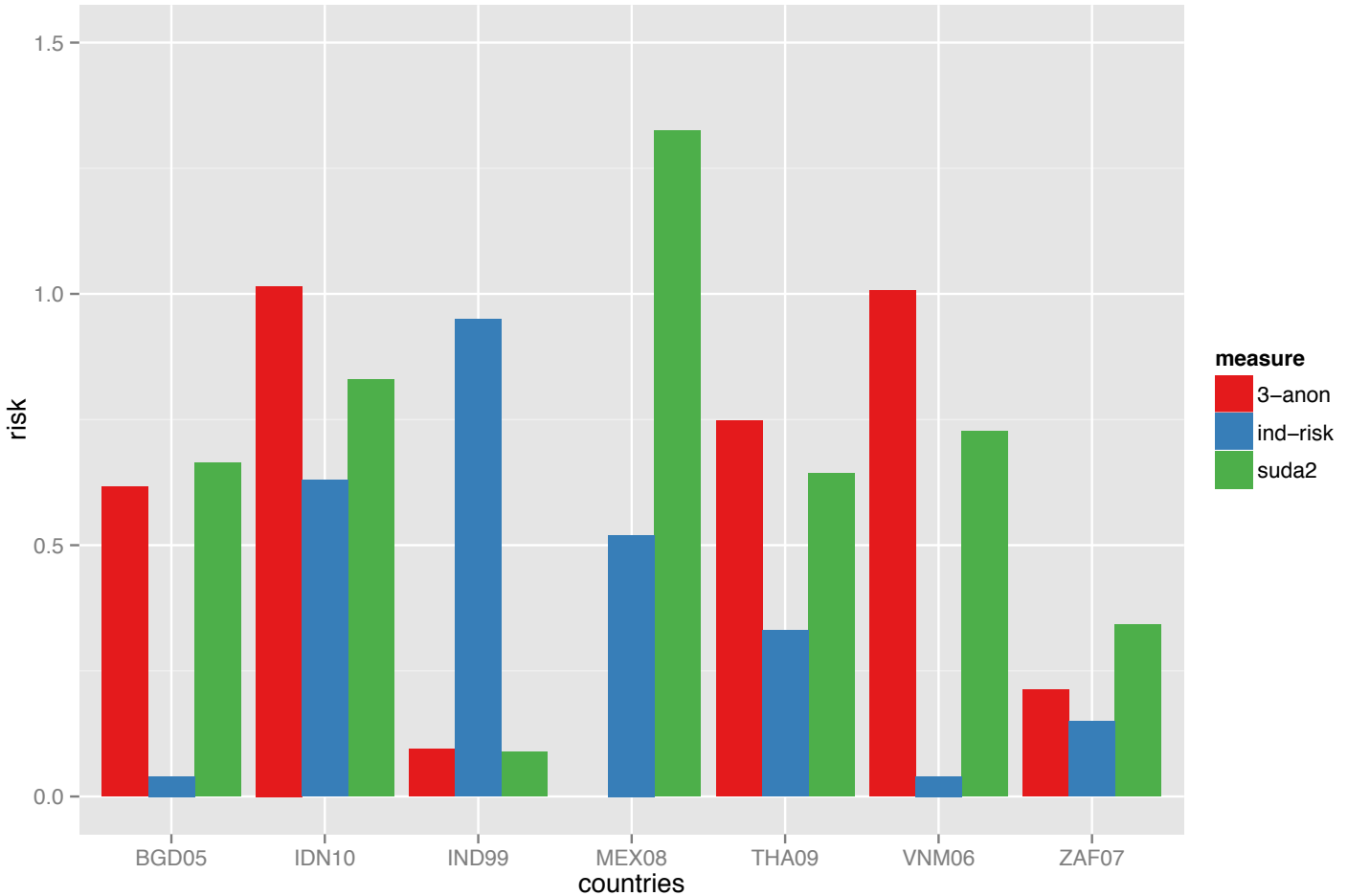
Multiple R-squared: 0.3338, Adjusted R-squared: 0.3336

Utility

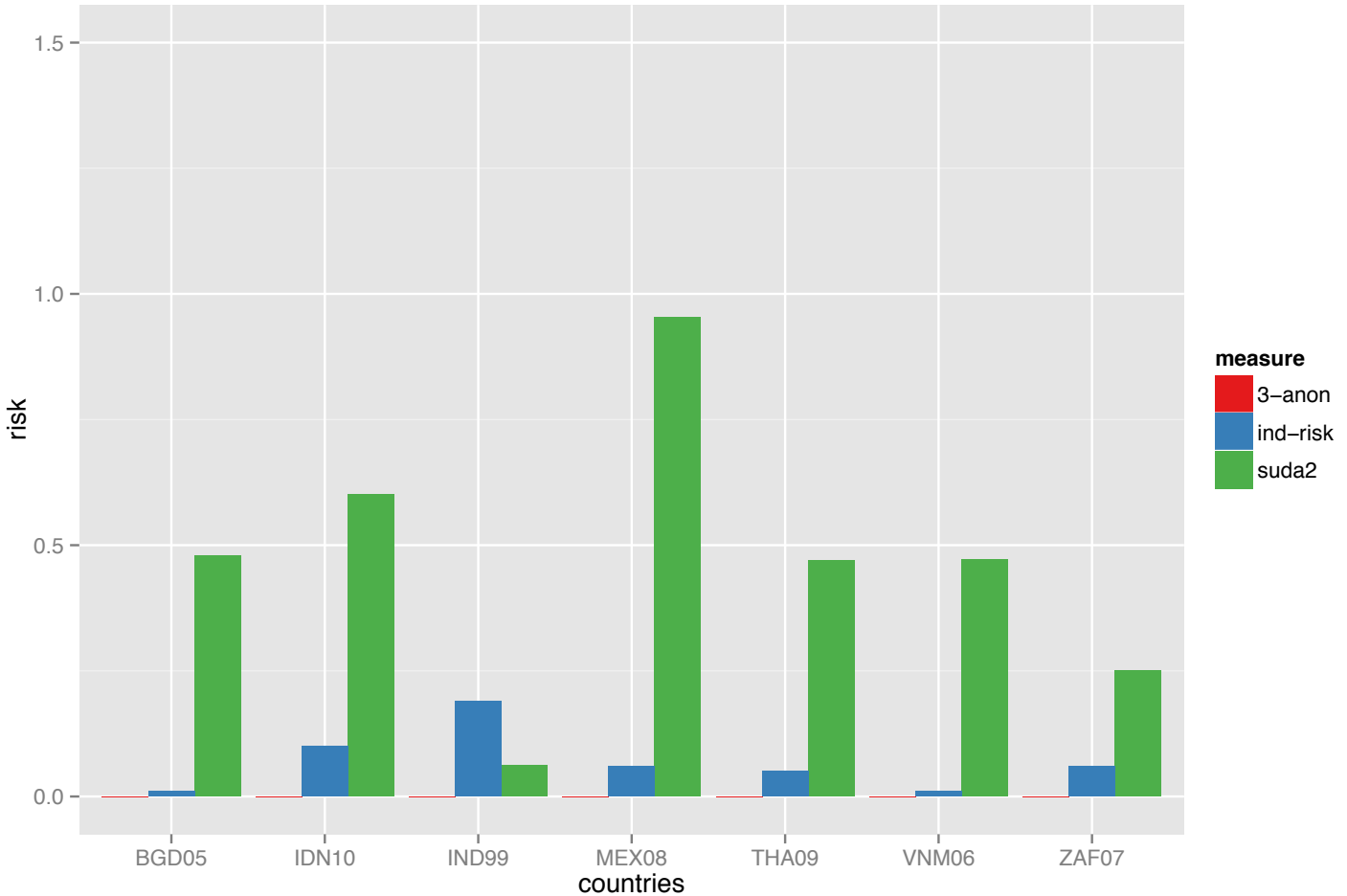
```
log(pcc) ~ age + literacy + edulevel2 + computer + urb
           + gender + reg01, data=bgd05sdc, weights=wgt
```

```
                          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
```

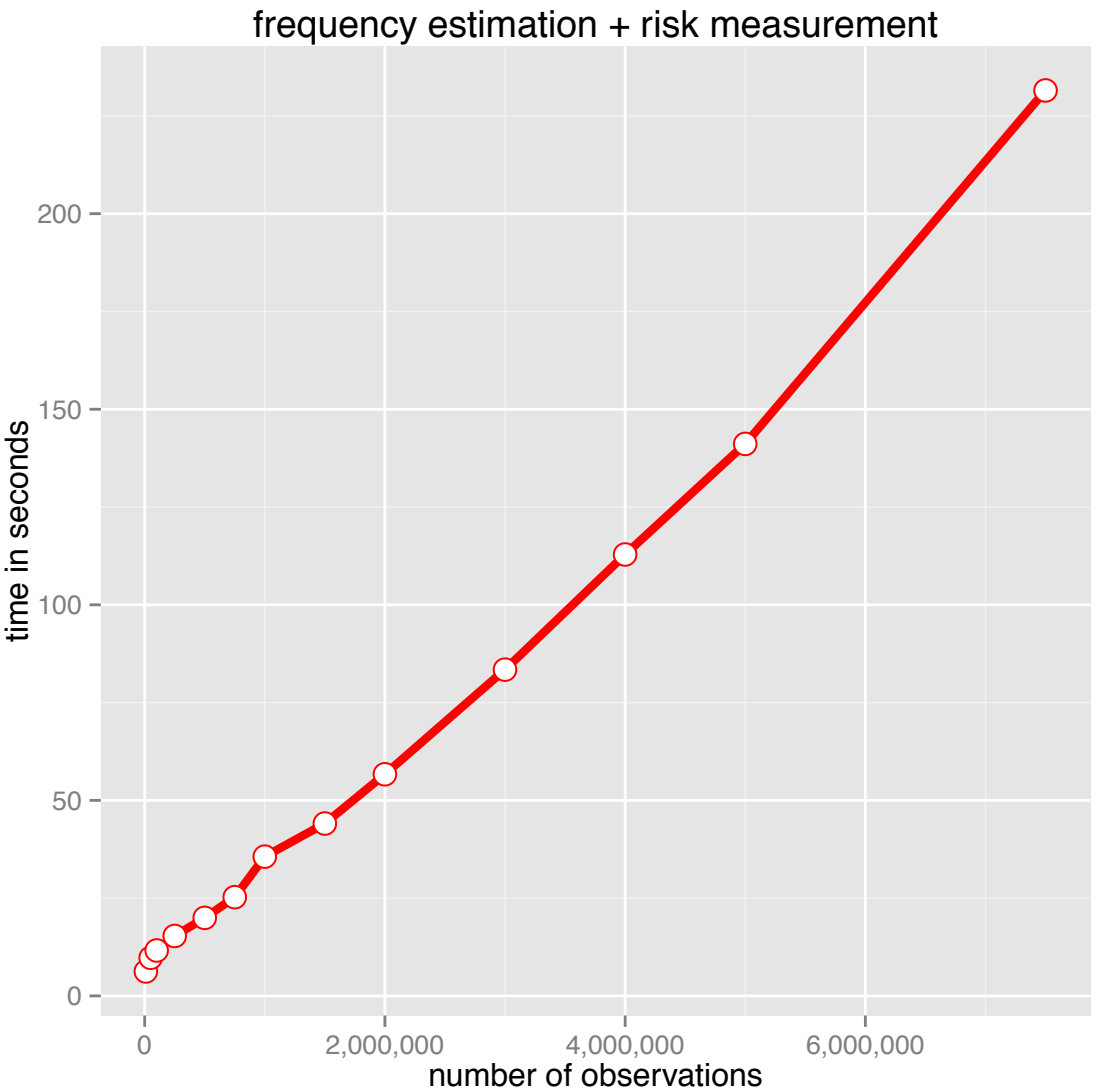
Application to Countries - Original Data



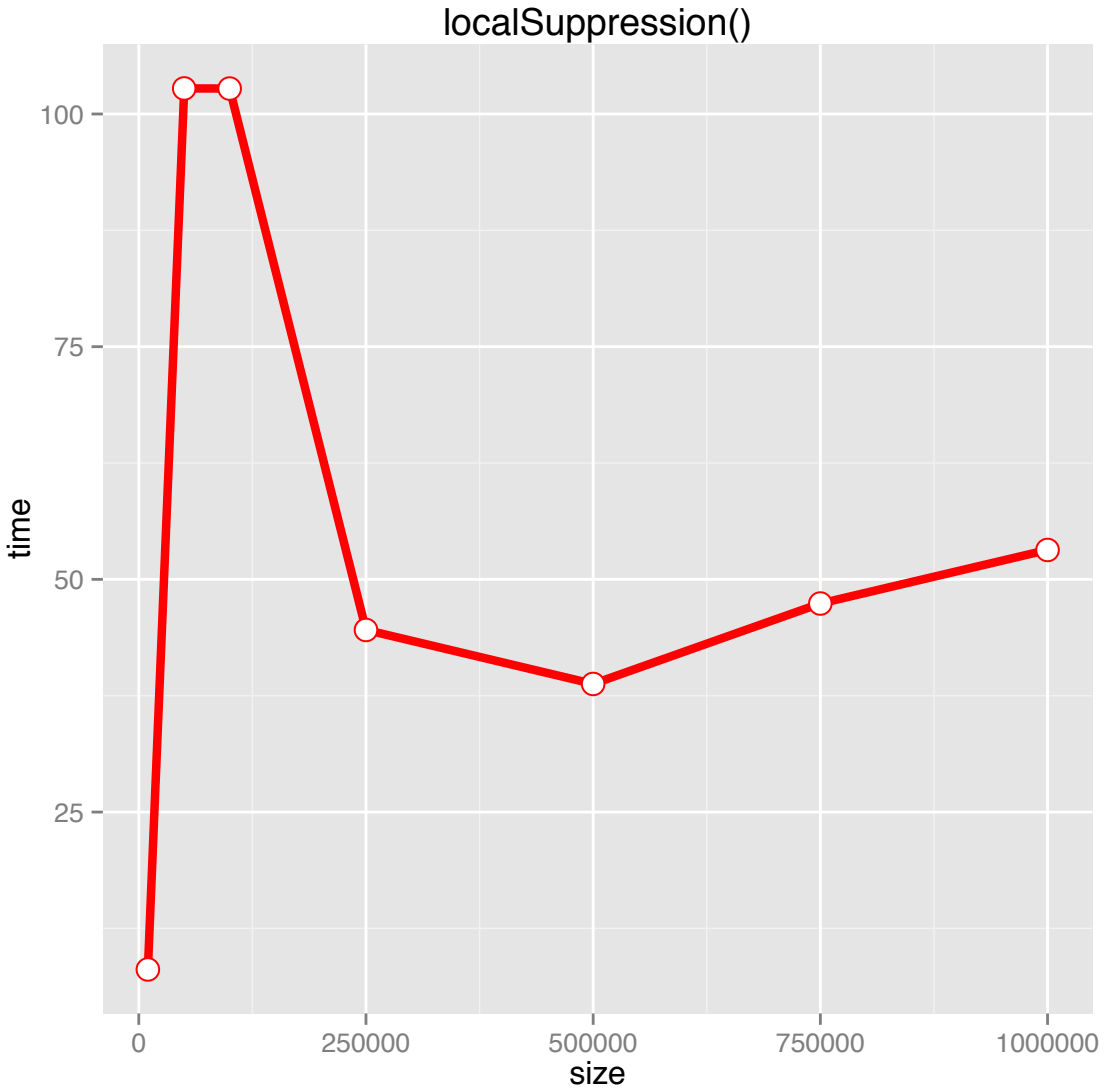
Application to Countries - Modified Data



Computational Costs



Computational Costs



Methods

Method | Software	μ-Argus 4.2	**sdcMicro** 1.0.0	**sdcMicro** > 4.3.0	**sdcMicroGUI** > 1.1.0	**IHSN**
frequency counts	✔	(✔)	✔	✔	✔
individual risk	✔	(✔)	✔	✔	✔
individual risk on households	✔		✔	✔	✔
l-diversity			✔	✔	✔
suda2			✔		✔
global risk	✔		✔	✔	✔
global risk with log-lin mod.			✔		
recoding	✔	(✔)	✔	✔	(✔)
local suppression	(✔)	(✔)	✔	✔	(✔)
swapping	(✔)	(✔)	✔		✔
pram	✔	✔	✔	✔	✔
adding correlated noise		✔	✔	✔	✔
microaggregation	✔	(✔)	✔	✔	✔
shuffling			✔	✔	
utility measures	(✔)	✔	✔	✔	
GUI	(✔)			✔	
CLI		✔	✔		✔
missing values	✔		✔	✔	✔
cluster designs	✔		✔	✔	✔
large data			✔	✔	(✔)
reporting	✔		✔	✔	
platform independent		✔	✔	✔	✔
free and open-source		✔	✔	✔	✔

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 ...