sdcMicro

A new flexible R-package for the generation of anonymised microdata:

Design issues and new methods

Matthias Templ

Manchester, Dezember 18, 2007

Outline of the talk

(1) Motives for the development of a new software for microdata protection.

(2) Requirements

(3) Package design and implementation

to avoid: 

(4) Example/Case study: Frequency counts, base individual risk (with no time for details...)

(6) Future developments

(1) Motivation

Perturbation of data is often the only possible way to provide data to researchers, because

- **Remote Access** to apply models without seeing the data is not in my mind (!).
- **Remote Access** with seeing synthetic data with the possibility of applying models on original data is not a good choice either, especially not when working with real complex data because outliers may disturb the analysis on original data.
- **Remote Access** with seeing the data is not possible in many countries nor for many data sources due to legal reasons.

In addition to that, **perturbation** of data does imply lower costs, because it can be achieved with less resources, also and the structure of the perturbed data is (approxiomately) the same as the one of original data with low information loss.

(2) Requirements for a SDC-software

- Powerful data import and export facilities
  - from free format (and fixed format), SPSS, SAS, STATA, Minitab, DBF, MySQL, Excel (some users insist on it), . . . and vise verca.
- Powerful CLI
  - If users wish to have a GUI, the software has to run in full functionality in CLI with a GUI on top of it.
- Reproducibility
  - It should be possible to reproduce all the results very easily (not by mouse clicking). Think of random number generators, script editors, . . .
- Cross platform portability
  - Software should run independently on all common (and exotic) platforms, by "platform" means the computer hardware and the operating system which run the application software
- Graphical excellence and implemented statistical methods
  - Microdata protection can also be exercised in an explorative way. Graphical tools and statistical methods are useful **during** the microdata protection process. If the software for SDC is implemented in a statistical software we are able to use some of the implemented statistical procedures (we are not reinventing the wheel here).

(2) Requirements for of a SDC-software

- Batch mode
  - It may be handy to run the software for SDC from another software via batch mode. Users demand this...
- Powerful programming language
  - Developing software for statistics is much easier (and more powerful) with an object oriented programming language (because we need to interact during the development in an effective way).
- Open source, free and flexible
  - It should be possible for everybody to contribute to the package's development or to extent the existing packages. So, the package has to be very flexible. But what is the disadvantage for me? ...

$\longrightarrow$ R does apply to all points above plus more (dynamical reports, ...).

Open source code, drawbacks

You can see my written code

$\longrightarrow$ you can see my possible mistakes

$\longrightarrow$ therefore, you may support a different career of mine ...

Locate a Prison
- Wandsworth
- Regime
- Visiting Information
- Employment Opportunities

How a Prison Operates

How Prisons are Regulated

Prison Service News (Magazine)

Contracted Out Prisons

Print this page

Email this page

Wandsworth

Wandsworth is a large prison in South West London, with a separate vulnerable prisoner unit. It is currently able to hold 1416 prisoners and is the largest prison in London. Alongside Liverpool, which is of similar size, it is one of the largest prisons in Western Europe.

The prison was built in 1851, and the residential areas remain in the original buildings. There has been extensive refurbishment and modernisation of the wings, including in-cell sanitation, privacy screens for cells occupied by more than one prisoner and the more recent installation of in-cell electricity.

Address:
PO Box 757
Heathfield Road
Wandsworth
London
SW18 3HS

Tel: 020 8588 4000
Fax: 020 8588 4001

Governor: Ian Mulholland

Operational Capacity: 1416 as of 30th June 2005

Accommodation: The Heathfield Centre (main prison) has five wings each with four landings, used as follows:

- A wing – Long term prisoners
- B Wing – 'Drug-free' Wing, Voluntary Testing Unit, CARATs (counselling, assessment, referral, advice and through care)
- C Wing – Induction Wing and Resettlement Unit
- D wing – Assessment and allocation, RAPt (Rehabilitation of Addicted Prisoners Trust 12 step programme)
- E wing – Pre-release and immigration detainees & Care and Separation unit
- The Onslow Centre (vulnerable prisoners unit) has 3 wings and holds approx. 330 prisoners.
- All wings have in-cell sanitation and in-cell electricity is currently being installed throughout the prison.

Reception Criteria: Normal reception arrangements: Wandsworth is a local prison, it accepts all suitable prisoners from courts in its catchment area.

(3) Package design

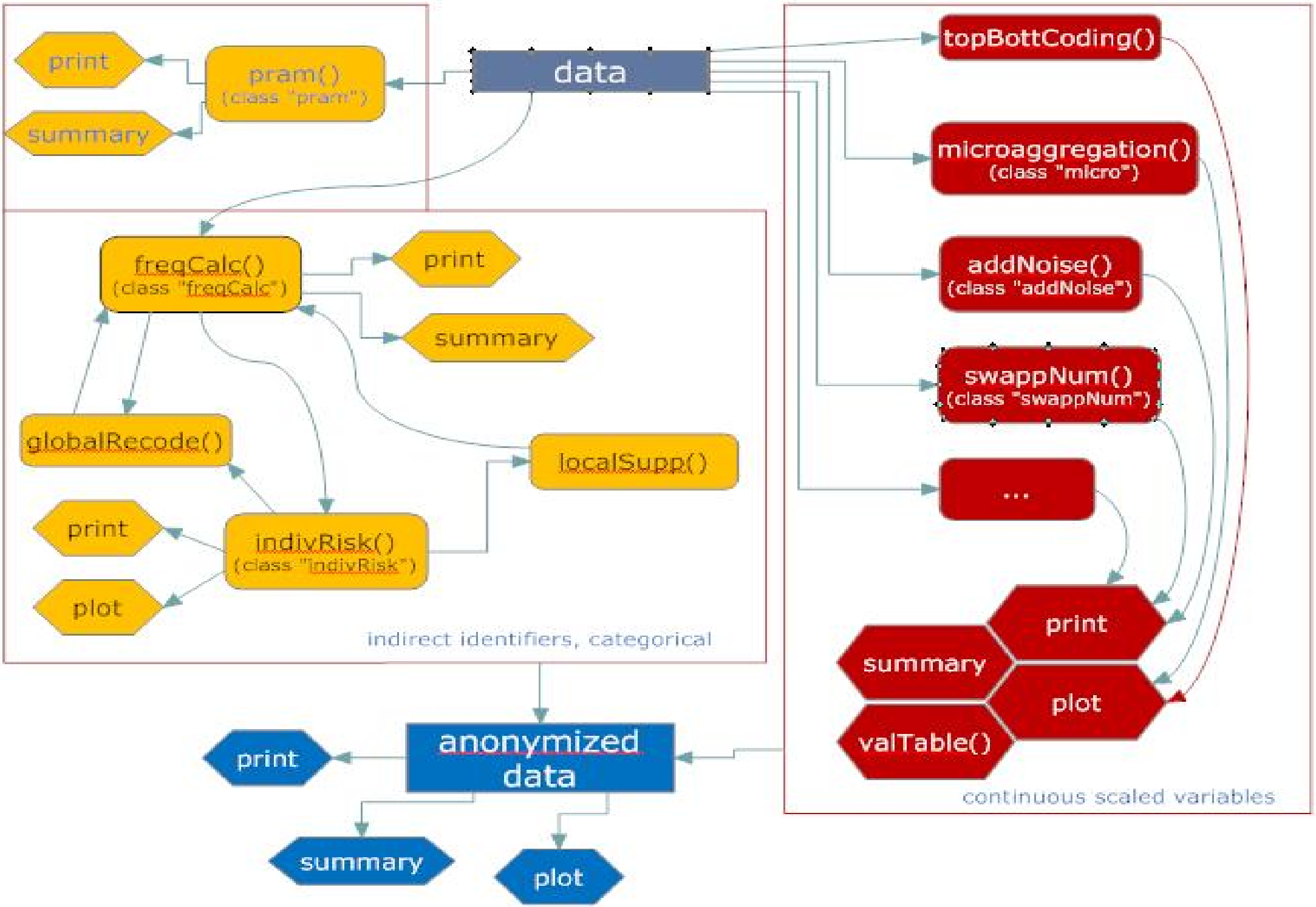
- In  everything is an object and every object is related to a specific class. One's own classes can be defined.
- The class of an object determines how it will be treated.
  - This class mechanism is extensively used in package *sdcMicro*.
- Nearly all functions in *sdcMmicro* produce objects from a certain class. Print, summary and plot methods are provided for objects from each of these classes.
  - `plot(ir1)` will plot, for example, a completely different plot as `plot(fc1)`, assuming that objects `ir1` and `fc1` are objects from different classes.
- This allows an easy-to-handle use of the package for every user.
  - print, summary and plot are names easy to remember.

(3) Package design, cont.

- Differnt methods can be tried out and their results can be compared easily.
  - Via comparison plots, measures of information loss.
- No tedious metadata management needs to be carried out by the user.
- Online documentation is included in the package.
  - Various examples are included for each of the functions which can be easily executed.
- Package is available on CRAN
  - http://cran.at.r-project.org/src/contrib/Descriptions/sdcMicro.html

Help pages for package `sdcMicro' version 2.1.0

sdcMicro-package	Statistical Disclosure Control (SDC) for the generation of protected microdata for researchers and for public use.
addNoise	Adding noise for the perturbation of data
casc1	Small Artificial Data set
CASCrefmicrodata	Census data set
dataGen	Fast generation of synthetic data
dRisk	overal disclosure risk
dUtility	data utility
EIA	EIA data set
francdat	data from the casc project
free1	Demo data set from mu-Argus
freqCalc	Frequencies calculation for risk estimation
globalRecode	Global Recoding
indivRisk	Individual Risk computation
localSupp	Local Suppression
microaggregation	Microaggregation
microData	microData
plot.indivRisk	plot method for indivRisk objects
plotMicro	Comparison plots
pram	Post RAndomisation Method (PRAM)
print.freqCalc	Print method for objects from class freqCalc
print.indivRisk	Print method for objects from class indivRisk
print.micro	Print method for objects from class micro
print.pram	Print method for objects from class pram
sdcMicro	Statistical Disclosure Control (SDC) for the generation of protected microdata for researchers and for public use.
summary.freqCalc	Summary method for objects from class freqCalc
summary.micro	Summary method for objects from class micro
summary.pram	Summary method for objects from class pram
swappNum	Rank Swapping
Tarragona	Tarragona data set
topBotCoding	Top and Bottom Coding
valTable	Comparison of different microaggregation methods


print
pram()
(class "pram")
summary
data
topBottCoding()
microaggregation()
(class "micro")
addNoise()
(class "addNoise")
swappNum()
(class "swappNum")
...
freqCalc()
(class "freqCalc")
print
summary
globalRecode()
localSupp()
print
indivRisk()
(class "indivRisk")
plot
indirect identifiers, categorical
print
summary
plot
valTable()
continuous scaled variables
anonymized data
print
summary
plot

(4) Example: frequency counts

Load the package and data after installation of *sdcMicro* from CRAN. The frequency counts are calculated as described in Capobianchi et al. (2001). R/C++ interface for fast calculation.

```
> library(sdcMicro)
> data(francdat)
> francdat
```

```
  Num1 Key1 Num2 Key2 Key3 Key4 Num3     w
1 0.30    1 0.40    2    5    1    4  18.0
2 0.12    1 0.22    2    1    1   22  45.5
3 0.18    1 0.80    2    1    1    8  39.0
4 1.90    3 9.00    3    1    5   91  17.0
5 1.00    4 1.30    3    1    4   13 541.0
6 1.00    4 1.40    3    1    1   14   8.0
7 0.10    6 0.01    2    1    5    1   5.0
8 0.15    1 0.50    2    5    1    5  92.0
```

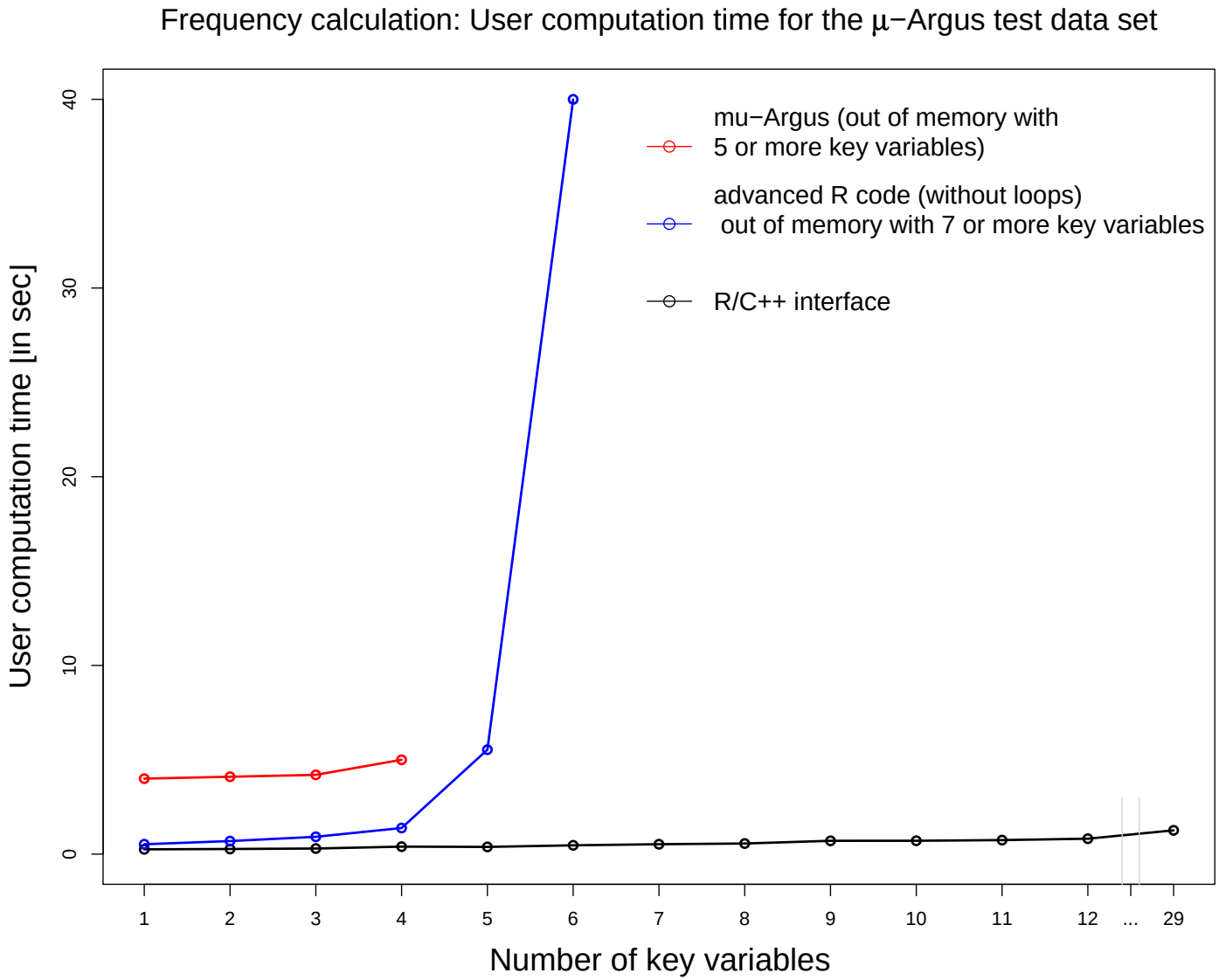
```
> f <- freqCalc(francdat, keyVars = c(2, 4, 5, 6), w = 8)
> class(f)
```

```
[1] "freqCalc"
```

(4) Example: frequency counts

```
> methods(class = freqCalc)

[1] print.freqCalc   summary.freqCalc

> f

 --------------------------
4 observation with fk=1
4 observation with fk=2
 --------------------------

> names(f)

[1] "freqCalc" "keyVars"  "w"        "indexG"   "fk"       "Fk"       "n1"
[8] "n2"

> f$Fk

[1] 110.0  84.5  84.5  17.0 541.0   8.0   5.0 110.0
```

(4) Example: frequency counts



(4) Example: individual risk

Let us calculate the individual risk for a part of the μ-Argus test data set.

```
> data(free1)
> f <- freqCalc(free1, keyVars = 1:3, w = 30)
> ind <- indivRisk(f)
> class(ind)

[1] "indivRisk"

> methods(class = indivRisk)

[1] plot.indivRisk  print.indivRisk

> ind

 ----- individual risk -----
  method = approx, qual = 1
 --------------------------

> plot(ind)
```

Datei History Resize Windows

R Console

```
R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreite
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden kö

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

> library(sdcMicro)
Lade nötiges Paket: car
Lade nötiges Paket: robustbase
Lade nötiges Paket: tcltk
Loading Tcl/Tk interface ... done
Lade nötiges Paket: cluster

 sdcMicro version 2.0.1 is already loaded
> data(free1)
> plot(indivRisk(freqCalc(free1, keyVars=1:3, w=30)))
<Tcl>
>
```

Individual risk adjustments

Please, see at the plot the active graphik device in R

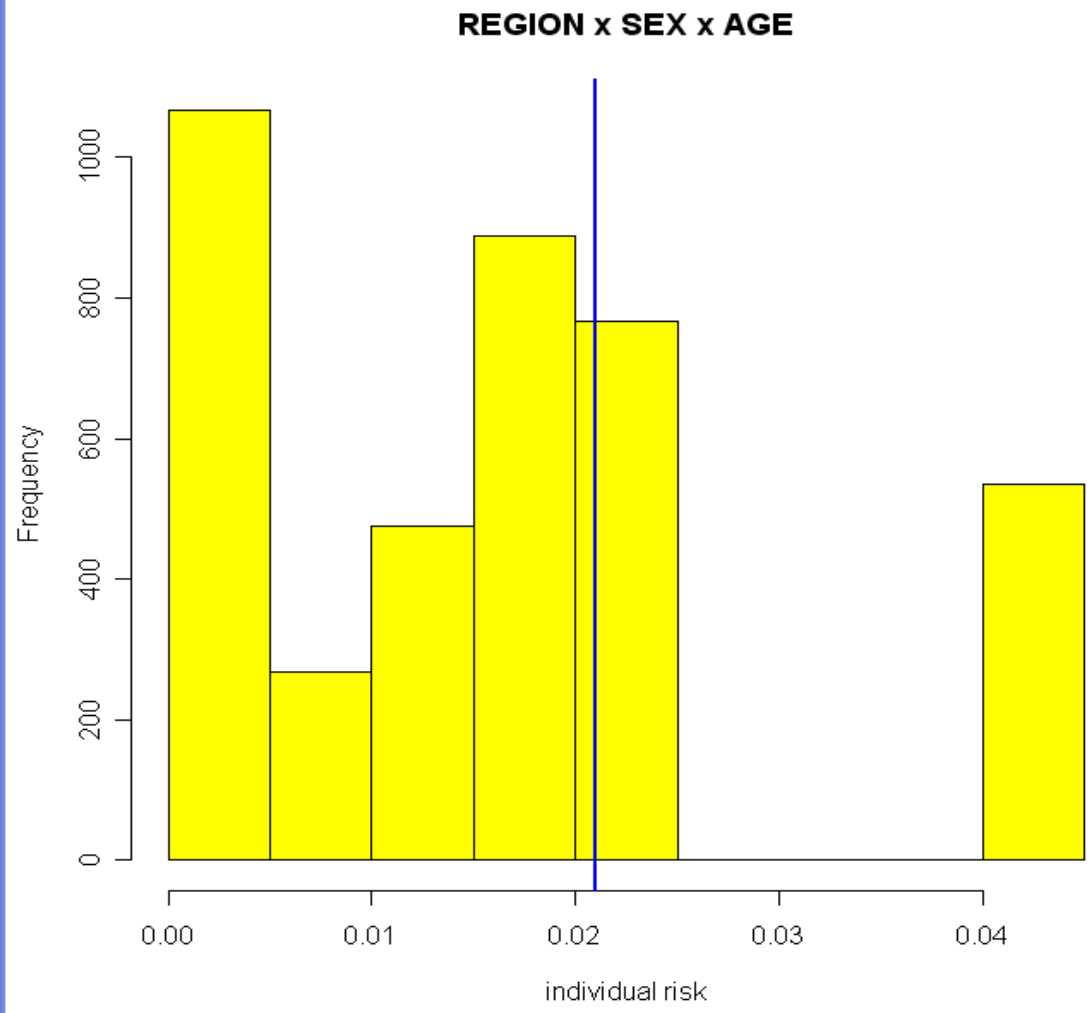
Individual risk threshold = 0.021

Re-identification rate = 1.58

Unsafe recods = 1300

ecdf Exit

R Graphics: Device 2 (ACTIVE)



Start TeXnicCenter - [is... isi.pdf - Adobe R... RGui Individual risk adj... Unbenannt - Paint DE 10:14

Datei History Resize Windows

R Console

```
R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreite
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden kö

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

> library(sdcMicro)
Lade nötiges Paket: car
Lade nötiges Paket: robustbase
Lade nötiges Paket: tcltk
Loading Tcl/Tk interface ... done
Lade nötiges Paket: cluster

 sdcMicro version 2.0.1 is already loaded
> data(free1)
> plot(indivRisk(freqCalc(free1, keyVars=1:3, w=30)))
<Tcl>
>
```

Individual risk adjustments

Please, see at the plot the active graphik device in R

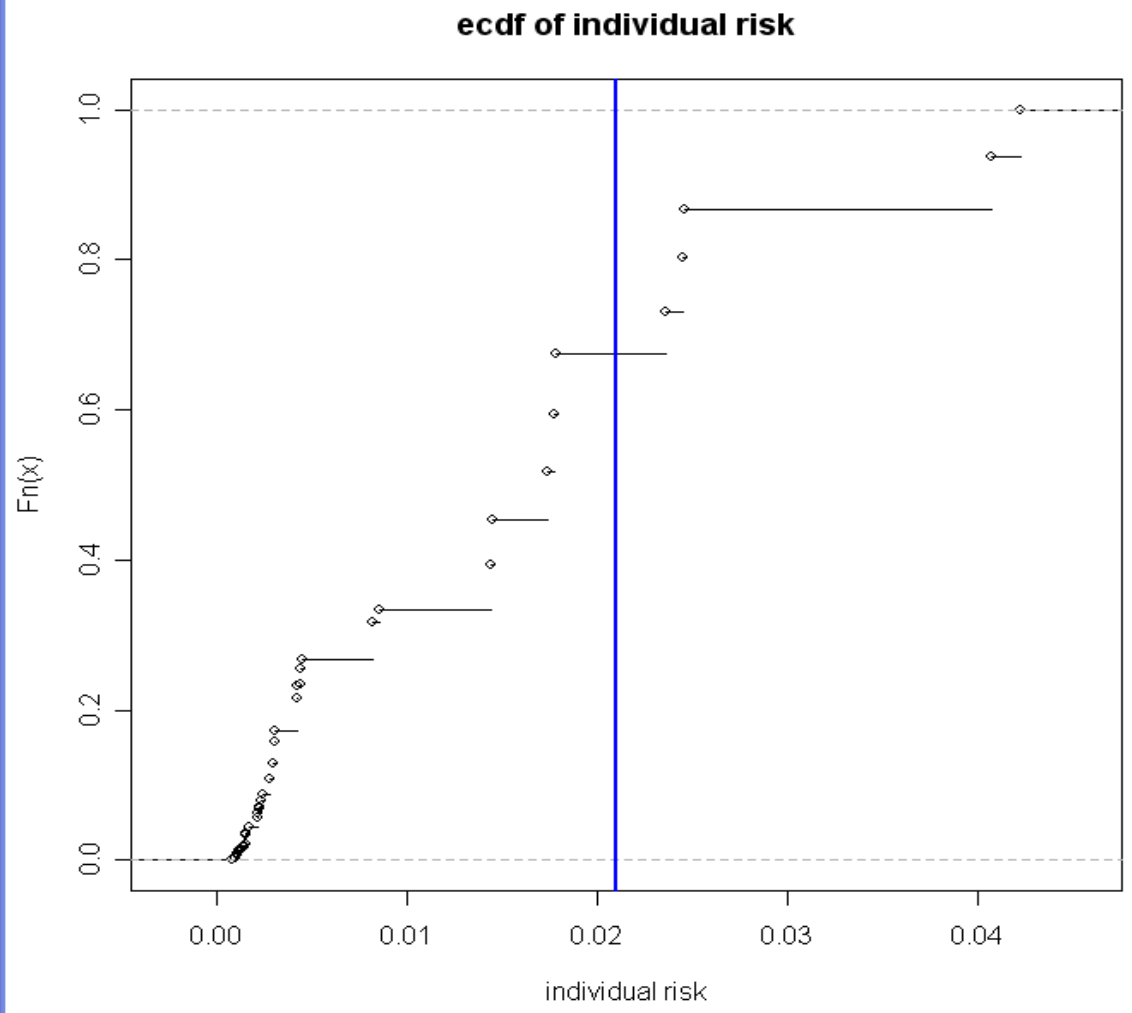
Individual risk threshold = 0.021

Re-identification rate = 1.58

Unsafe recods = 1300

ecdf Exit

R Graphics: Device 2 (ACTIVE)



Start TeXnicCenter - [is... isi.pdf - Adobe R... RGui Individual risk adj... Unbenannt - Paint DE 10:13

Measures of information loss

A broad range of well known (and new) measures of information loss are implemented for

- The comparison of original data and perturbed data in accordance with univariate and multivariate statistics.
- The use of numerous of comparison plots which are more powerful.
- But measures for disclosure risk and data utility are also provided.

And you can use the whole power of  to visualise and compare your results.

Measures of information loss

```
> g <- valTable(Tarragona, method=c("addNoise: correlated2", "swappNum",
                        "simple", "individual ranking", "pca", "clustpca",
                        "mdav", "rmdm2", "clustpppca" ) )
```

```
> xtable1(g, c(1, 3:7, 12)])
```

	method	amean	amedian	devvar	amad	acov	apcaload
1	addNoise: correlated2	0.09	3.20	0.18	5.89	0.09	8.97
2	swappNum	0.20	0.13	9.10	0.23	4.55	28.75
3	simple	0.00	3.50	3.62	1.11	1.81	20.69
4	individual ranking	0.00	0.02	13.71	0.03	6.85	19.45
5	pca	0.00	2.62	2.77	1.10	1.39	12.99
6	clustpca	0.00	2.34	2.69	1.11	1.35	11.72
7	mdav	0.00	4.18	3.44	1.84	1.72	8.76
8	rmdm2	0.00	1.51	1.76	0.58	0.88	12.00
9	clustpppca	0.00	3.64	2.79	1.55	1.40	10.69

Furture developments

- Imputation methods for the generation or perturbation of data
- Bernhard Meindl and I are working on `sdcTable` at the moment.

Conclusions

- The package has a lot of advantages for the effective use of SDC methods for the generation of anoymised microdata.
- Fast C++ code for time consuming calculation steps. *
- The package is used by Statistics Austria for the generation of anonymised microdata for five months now, and is already used by research organisations and companies.
- Contributions to this package are highly welcome.

*Thanks to Bernhard Meindl for the collaboration on the C stuff.