sdcMicro

- The R package **sdcMicro** is a well-known collection of microdata protection methods developed by Statistics Austria, which is already in use in several national statistics offices.
- sdcMicro has become one of the standard tools for microdata protection during the last five years.
- The IHSN is supporting the further development of sdcMicro and has partnered with its developers to perform the following tasks (next slide).

- Include in **sdcMicro** relevant methods available in the IHSN plug-ins
- Test **sdcMicro** on real datasets to calibrate its outputs and facilitate their interpretation
- Develop practical guidelines to support the use of a toolbox and help users navigate between methods and associated algorithms

SdcMicro already includes several popular methods for microdata anonymization; some of these methods can also be found in the IHSN C++ plug-ins.

- The overlapping methods have been tested and compared with their analogous implementation in **sdcMicro**.
- Three new methods (or improved implementations) have been included in sdcMicro: suda2 (i.e., finding minimal samples unique), rank swap (i.e., numerical rank swapping), and mdav (i.e., micro-aggregation).
- Since the C++ code contained specific class structures and required multiple and sometimes different header files to be included when compiling the code, the inclusion of these new methods into `R` has proved to be a complex task.

The following Figure 1 shows computation time efficiency gains between the old and new implementation of rank swapping in sdcMicro, based on 100 runs on a 10-dimensional dataset with varying numbers of observations.
Figure 1 computation time efficiency gains between old and new implementation of rank swapping algorithms
SDC

- Version 4.4.0 of the sdcMicro package is available on the Comprehensive R Archive Network (CRAN).
- Existing guidelines and a user guide for **sdcMicro** are being updated.
- A specific tutorial is being developed to show how to implement these concepts and algorithms on real datasets.

Statistical Disclosure Control

- This tutorial is being drafted with examples of the European Union SES dataset.
- The IHSN is promoting the adoption of sdcMicro and an associated guidelines toolbox for the creation of Public Use Files and Scientific Use Files.
- See our pages Software - Statistical Disclosure Control and Guidelines on Microdata Anonymization.

- Class `"sdcMicroObj"`
- Data shuffling and General Additive Data Perturbation.
- `suda2` Detecting Special Uniques

```
> data(testdata)
> sdc <- createSdcObj(testdata,
+ keyVars=c('urbrur','roof','walls','relat','sex'),
+ pramVars=c('water','electcon'),
+ numVars=c('expend','income','savings'), w='sampling_weigh
> fk=freq(sdc)
> Fk=freq(sdc,type="Fk")
```

```
> print(sdc)
Number of observations violating

-  2-anonymity:  26
-  3-anonymity:  52
---------------------------

Percentage of observations violating
-  2-anonymity:  0.57 %
-  3-anonymity:  1.14 %
```

```
> print(sdc,type="ls")

urbrur .. 0 [ 0 %]

roof .... 0 [ 0 %]

walls ... 0 [ 0 %]

relat ... 0 [ 0 %]

sex ..... 0 [ 0 %]
```

```
> print(sdc,type="recode")
Reported is the
number | mean size and | size of smallest category

-------------
urbrur .. 2 | 2290 | 646
-------------
roof .... 5 | 916 | 16
-------------
walls ... 3 | 1527 | 50
-------------
relat ... 9 | 509 | 1
-------------
sex ..... 2 | 2290 | 2284
```

```
> print(sdc,type="risk")

---------------------------
0 obs. with higher risk than the main part
Expected no. of re-identifications:
2.41 [ 0.05 %]
---------------------------
```

```
> print(sdc,type="numrisk")
Disclosure Risk is between:
[0% ; 100%] (current)
- Information Loss:
IL1: 0
- Difference Eigenvalues: 0 %
```

```
> print(sdc,type="pram")
PRAM has not been applied!
>
```