sdcApp – a shiny new GUI for sdcMicro

Bernhard Meindl\textsuperscript{1}, Alexander Kowarik\textsuperscript{1}, Matthias Templ\textsuperscript{2}, Matthew Welch\textsuperscript{3}, Thijs Benshop\textsuperscript{4}

\textsuperscript{1}Statistik Austria, \textsuperscript{2}ZHAW, \textsuperscript{3}WorldBank, \textsuperscript{4}Humboldt University

September, 2017
Introduction

- NSIs publish a wide range of high-quality statistical output
- Demand for microdata from researchers increases
- Aim: provide datasets for release that don’t allow users to link information to specific entities
- R-package sdcMicro provides the tools ⇒ sdcApp() provides a GUI - no R knowledge necessary
- github: https://cran.r-project.org/web/packages/sdcMicro
- developed within the sdcTools framework: https://github.com/sdcTools/sdcMicro

This work has been funded by the World Bank under the project "DEVELOPING AN R GRAPHIC USER INTERFACE FOR STATISTICAL DISCLOSURE CONTROL (SDC)", selection number 1216510.
let's try it out and start the shiny GUI

```r
> library(sdcMicro)
> sdcApp()
```
First impressions

sdcApp

This graphical user interface of sdcMicro allows you to anonymize microdata even if you are not an expert in the R programming language. Detailed information on how to use this graphical user-interface (GUI) can be found in a tutorial (a so-called vignette) that is included in the sdcMicro package. The vignette is available from the CRAN website or by typing vignette("sdcApp", package="sdcMicro") into your R prompt.

For information on the support and development of the graphical user interface, please click here.

Getting started

To get started, you need to upload a file with microdata to the GUI. You can do so by clicking this button. Alternatively, you can upload a previously saved problem instance by clicking here.

Set storage path

Currently, all output, such as anonymized data, scripts and reports, will be saved to /Users/thijsbenschop.

You can change the default path, where all output from the GUI will be saved. You can change this path any time later as well by returning to this tab.

Enter a directory where any exported files (data, script, problem instances) should be saved to

e.g.: /Users/thijsbenschop

Stop the interface

By clicking Stop the GUI, you can stop the graphical user interface at any time during the anonymization process. If you have started the interface as x <- sdcApp(), x will contain the micro data and the sdc problem at the state just before stopping the GUI.

Restart the interface
Uploading Microdata

Select data source:
- Testdata/Tabular data
- R-dataset (.rda)
- SPSS-file (.sav)
- SAS-file (.sas786)
- CSV-file (.csv, .txt)
- STATA-file (.dta)

Uploading microdata
Load the dataset to be anonymized.

Select a test dataset or any object in your current workspace:
- testdata

Load data
Microdata have been uploaded

The loaded dataset is **Test data** and consists of 4580 observations and 15 variables. No variables were dropped because of all missing values.

<table>
<thead>
<tr>
<th></th>
<th>usbrur</th>
<th>roct</th>
<th>walls</th>
<th>water</th>
<th>electron</th>
<th>relat</th>
<th>sex</th>
<th>age</th>
<th>hhcivil</th>
<th>expend</th>
<th>income</th>
<th>savings</th>
<th>ori_hid</th>
<th>sampling_weight</th>
<th>household_weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>46</td>
<td>2</td>
<td>9892969</td>
<td>5780000</td>
<td>136258.5</td>
<td>1</td>
<td>108</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>41</td>
<td>2</td>
<td>2733858</td>
<td>2530000</td>
<td>279345</td>
<td>1</td>
<td>108</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>2652417</td>
<td>6028000</td>
<td>5495381</td>
<td>1</td>
<td>108</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1887384</td>
<td>7068000</td>
<td>8695862</td>
<td>1</td>
<td>108</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>52</td>
<td>1</td>
<td>6731247</td>
<td>9838000</td>
<td>203628.2</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>47</td>
<td>2</td>
<td>498576</td>
<td>5290000</td>
<td>1821268</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>6339509</td>
<td>2270000</td>
<td>8129160</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>1184874</td>
<td>8018000</td>
<td>8818464</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>3260930</td>
<td>207324</td>
<td>7043642</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>18</td>
<td>1</td>
<td>3434769</td>
<td>4418000</td>
<td>4783134</td>
<td>2</td>
<td>108</td>
<td>16.6666666666667</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>65</td>
<td>2</td>
<td>7180354</td>
<td>5558000</td>
<td>7942221</td>
<td>3</td>
<td>108</td>
<td>33.3333333333333</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>2</td>
<td>5517434</td>
<td>4120000</td>
<td>4318171</td>
<td>3</td>
<td>108</td>
<td>33.3333333333333</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>4680282</td>
<td>9168000</td>
<td>2688967</td>
<td>3</td>
<td>108</td>
<td>33.3333333333333</td>
<td></td>
</tr>
</tbody>
</table>
Anonymize tab to setup an SDC problem
Suppressing values in high-risk obs

This method allows to suppress (set to NA) values in the selected key variables for records that have an individual risk higher than the specified threshold.

Select key variable for suppression

Threshold for individual risk

roof

Suppress 4580 values with high risk in "roof"
### Risk and Utility tab

#### Tabular representation of original and modified data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Original data</th>
<th>Modified data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>Freq</td>
</tr>
<tr>
<td>roof</td>
<td>814</td>
<td>814</td>
</tr>
<tr>
<td>2</td>
<td>6397</td>
<td>6397</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>Sum 4580</td>
</tr>
</tbody>
</table>

#### Variable selection

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Type</th>
<th>Suppressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>roof</td>
<td>cat. key variable</td>
<td>0</td>
</tr>
<tr>
<td>income</td>
<td>num. key variable</td>
<td>0</td>
</tr>
<tr>
<td>savings</td>
<td>num. key variable</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Additional parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of records</td>
<td>4580</td>
</tr>
<tr>
<td>alpha</td>
<td>1</td>
</tr>
<tr>
<td>random seed</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Anonymization steps

- Recoded "roof": "6", "9" to "6_9"
Reproducibility

View the current generated script

Browse and download the script used to generate your results. These can be used later as a reminder of what you did or entered into R from command-line to reproduce results.

```r
require(sdcMicro)
inputdata <- readMicrodata(path="testdata", type="rdf", convertCharToFac=FALSE, drop_all_missings=FALSE)
inputdataAll <- inputdata

inputdata <- varToFactor(obj=inputdata, var="age")
## Set up sdcMicro object
sdcObj <- createSdcObj(data=inputdata,
                   keyVars=c("roof"),
                   numVar=c("income","savings"),
                   weightVar=NULL,
                   hhId==NULL,
                   strataVar=NULL,
                   promVar= NULL,
                   excludeVars= NULL,
                   seed=0,
                   randomizeRecords=FALSE,
                   alpha=1)

## Store name of uploaded file
opts <- get.sdcMicroObj(sdcObj, type="options")
optssfilename <- "testdata"
sdcObj <- set.sdcMicroObj(sdcObj, type="options", input=list(opts))

## Recode variable
sdcObj <- groupAndRename(obj=sdcObj, var="roof", before=c("6","9"), after=c("6_9"), addNA=FALSE)
```
Undo-functionality

Undo last step
Clicking the button below will remove (if possible) the following anonymization step!

Recoded “roof”: “6”, “9” to “6_9”

Undo last step

Save and retrieve current state
The undo button can only be used to go one step back. For experimenting with SDC methods, parameters and settings, it can be useful to save a certain state before starting to experiment with different SDC methods and, if the result is not satisfactory, revert to the saved state. Here you can save the current state and, if necessary, reload this state. Reloading undoes any methods applied to the data since saving the last state, but restores any methods applied before the saving. It is also possible to save several states, as they are saved on disk.

Note: This feature is GUI-only and cannot be reproduced from the command-line version.

Save current state
Click here to save the current state with all relevant data and code for reverting to this state later. This can also be used to save the current state and continue working on this SDC problem at a later point in time.

Save current state

Revert to saved state
Here you can load a previously saved state. The file must be an .rdata file. See above for the path where you saved the last state. Please note that uploading a previously saved state overwrites all current results and results into a loss of any unsaved changes.

Select previously exported sdcProblem (.rdata)

Browse... No file selected
Summary

- **sdcMicro** is actively developed
- **sdcApp()** allows non-expert R-users to create safe datasets
- already successfully used in practical sdc trainings
- additional information is available by hovering over questionsmark-symbols scattered throughout the GUI
- Feedback ([https://github.com/sdcTools/sdcMicro/issues](https://github.com/sdcTools/sdcMicro/issues)) or contributions are very welcome