



# Statistical Disclosure Control: A Graphic Interface for sdcMicro

Matthew Welch



# Background

- Application of many anonymization methods is complex and requires knowledge of the methods and access to suitable tools for implementation
- For users comfortable with using R, the package sdcMicro provides a tool for the application of methods commonly described in the literature on disclosure control.
- For users not familiar with R, a Graphic User Interface (GUI) for the sdcMicro package may be useful.
- To provide a GUI environment for the non-R user, a Shiny application called sdcApp has been added to the sdcMicro package.
- sdcMicro is open source and available in the [CRAN repositories](#) and on [GitHub](#).

# sdcMicro Functionality

- [Developed by Alexander Kowarik, Matthias Templ and Bernhard Meindl, R-Package “sdcMicro” \(2017\)](#)
- anonymization methods, risk/utility measures, and data manipulation functions
- anonymization methods include, global recoding, local suppression, post-randomization (PRAM), top- and bottom-coding, microaggregation, noise addition and rank swapping.
- for estimating disclosure risk - k-anonymity, individual and global risk, l-diversity, SUDA scores and proximity measures.
- for information loss measures like differences in eigenvalues and tabulations



# A Shiny interface for sdcMicro

- World Bank\IHSN funded the development – funding DFID to IHSN
- Aim was to provide a GUI that removes the need for proficiency in R, but still allows access to all the features of sdcMicro.
- The Shiny application, sdcApp, implements all the main anonymization methods available in the sdcMicro package
- In addition, sdcApp offers a comprehensive set of risk and utility measures.
- This includes functions to measure, visualize and compare risk and utility throughout the anonymization process.



# A Shiny interface for sdcMicro

- a number of features are brought in from other R packages
- These make measuring utility, visualizing and exploring the data and changes made in the SDC process easier
- sdcApp also helps users by producing reports on the methods used and saves the underlying R code to ensure reproducibility.
- For users of other statistical packages, sdcApp supports importing and exporting microdata in several formats (STATA, SAS, SPSS, CSV, R).

# Installing and launching the GUI

- To use the application, the user needs to install the latest version of the R software from the CRAN website: <https://cran.r-project.org>
- After installing R, the user installs the `sdcMicro` package and dependencies
- After installing `sdcMicro`, the package `sdcMicro` needs to be loaded with the command `library(sdcMicro)`. `sdcApp` is then launched with the command `sdcApp()`
- The application launches in the default web browser of your system.
- The user can interact with the application by using control inputs such as buttons, drop-down menus, sliders, radio buttons or text input. No further interaction with the R console is required.

# Layout of the GUI

- sdcApp consists of seven tabs that can be navigated using the top navigation bar.
- About/Help -> Help and general settings
- Microdata -> Load and prepare dataset
- Anonymize -> Anonymization methods
- Risk/Utility -> Risk and utility measures
- Export Data -> Export data, reports
- Reproducibility -> Generate R script
- Undo -> Undo steps in anonymization process

# About and settings

sdcMicro GUI

About/Help

Microdata

Anonymize

Risk/Utility

Export Data

Reproducibility

Undo

## 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 on [GitHub pages](#) and via the [CRAN](#) website. The vignette can also be viewed offline by typing `vignette("sdcMicro", 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 `/volumes/transcend/world bank/`.

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





# Load the data

**sdcmicro GUI**    About/Help    **Microdata**    Anonymize    Risk/Utility    Export Data    Reproducibility    Undo

**Select data source**

- Testdata/internal data
- R-dataset (.rdata)
- SPSS-file (.sav)
- SAS-file (.sas7dat)
- CSV-file (.csv, .txt)
- STATA-file (.dta)**

## Uploading microdata

Load the dataset to be anonymized.

### Set additional options for the data import

Convert string variables (character vectors) to factor variables?  TRUE  FALSE

Drop variables with only missing values (NA)?  TRUE  FALSE

Note: the selected file is loaded immediately upon selecting. Set the above options before selecting the file.

Select file (allowed types are '.dta')

No file selected

# Explore and prepare

**sdcMicro GUI**    About/Help    Microdata    Anonymize    Risk/Utility    Export Data    Reproducibility    Undo

**Loaded microdata**

The loaded dataset is `case_1_data_lab.dta` and consists of **10574** observations and **66** variables. **2** variable(s) was/were dropped because of all missing values: **ETHNICITY**  
**LANGUAGE**

Show  entries    ThirdFilter:

REGION	DIST	URBRUR	WGTHH	WGTPOP	IDH	IDP	HHSIZE	GENDER	REL
Region 4	411	Urban	425.038421630859	2975.26904296875	1	1	7	Male	Head
Region 4	411	Urban	425.038421630859	2975.26904296875	1	2	7	Female	Spouse
Region 4	411	Urban	425.038421630859	2975.26904296875	1	3	7	Male	Child
Region 4	411	Urban	425.038421630859	2975.26904296875	1	4	7	Male	Child

Showing 1 to 20 of 10,574 entries

Previous **1** 2 3 4 5 ... 529 Next

**What do you want to do?**

- Display microdata
- Explore variables
- Reset variables
- Use subset of microdata
- Convert numeric to factor
- Convert variables to numeric
- Modify factor variable
- Create a stratification variable
- Set specific values to NA
- Hierarchical data

# Explore and prepare

sdcmicro GUI    About/Help    Microdata    Anonymize    Risk/Utility    Export Data    Reproducibility    Undo

What do you want to do?

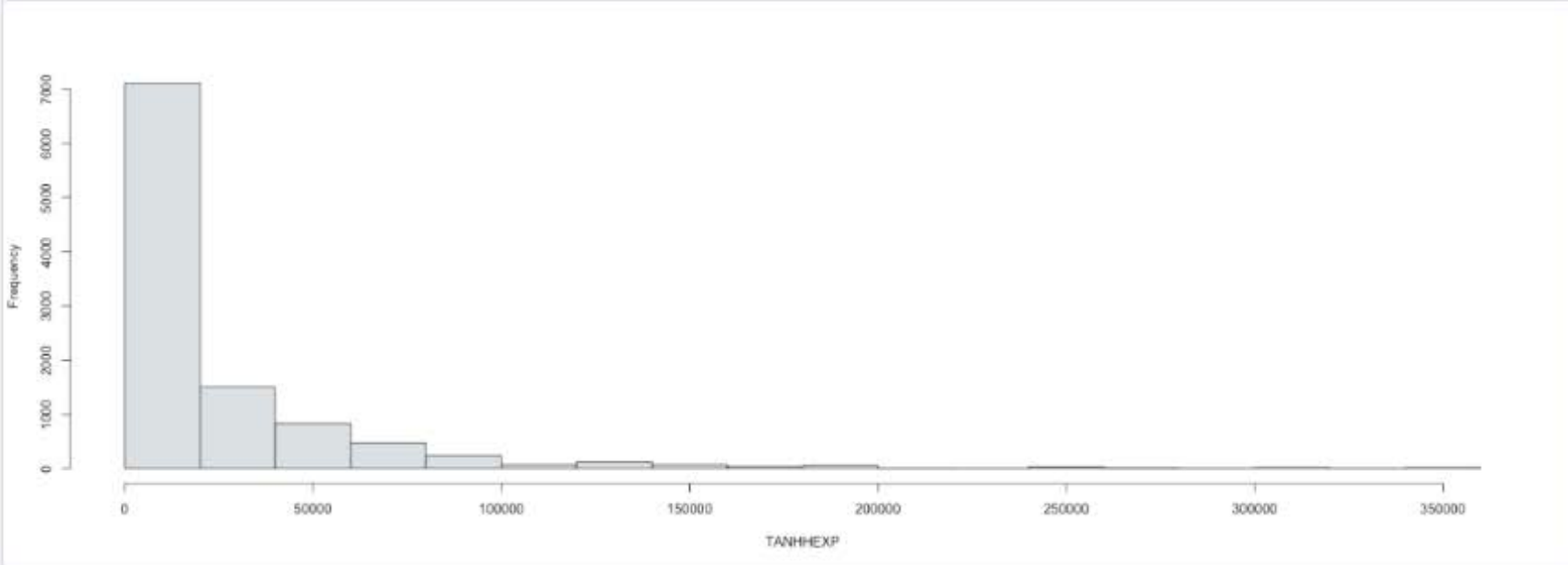
- Display microdata
- Explore variables**
- Reset variables
- Use subset of microdata
- Convert numeric to factor
- Convert variables to numeric
- Modify factor variable
- Create a stratification variable
- Set specific values to NA
- Hierarchical data

Reset input data

## Explore variables in original data

Here you can view tabulations, summary statistics and graphic representations of variables and pairs of variables to explore the original data.

Choose a variable: **TANHHEXP (numeric)**    Choose a second variable (optional): **none**



The histogram displays the frequency distribution of the variable TANHHEXP. The x-axis represents the value of TANHHEXP, ranging from 0 to 350,000 with major ticks every 50,000. The y-axis represents the frequency, ranging from 0 to 7,000 with major ticks every 1,000. The distribution is highly right-skewed, with the highest frequency (approximately 7,000) occurring in the first bin (0 to 25,000). The frequency drops significantly for subsequent bins, with values around 1,500, 800, 500, and 300 for the next four bins. The distribution continues to be sparse with very low frequencies up to 350,000.

Min	Q5	Q25	Median	Mean	Q75	Q95	Max
498.00	739.00	15547.00	17293.00	28560.00	29722.00	86319.00	353230.00

Variable **TANHHEXP** has 0 ( 0.00% ) missing values.

# Selecting key variables

sdcMicro GUI

About/Help

Microdata

Anonymize

Risk/Utility

Export Data

Reproducibility

Undo

## Anonymize

Select variables and set parameters to create the SDC problem.

### Select variables i

Variable name	Type	Key variables			Weight	Hierarchical identifier	PRAM	Dele
REGION	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
DIST	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
URBRUR	factor	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WGTHH	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WGTPOP	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
IDH	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
IDP	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HHSIZE	numeric	<input checked="" type="radio"/> No	<input type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
GENDER	factor	<input type="radio"/> No	<input checked="" type="radio"/> Cat.	<input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### Set additional parameters

Parameter 'alpha' i



Parameter 'seed' i



### Explore variables

REGION (factor) v



# Summary view of the SDC problem

**sdcMicro GUI**    About/Help    Microdata    **Anonymize**    Risk/Utility    Export Data    Reproducibility    Undo

---

**View/Analyze existing sdcProblem**

- Show summary
- Explore variables
- Add linked variables
- Create new IDs

**Anonymize categorical variables**

- Recoding
- k-Anonymity
- PRAM (simple)
- PRAM (expert)
- Supress values with high risks

**Anonymize numerical variables**

---

## Summary of dataset and variable selection

The loaded dataset consists of **10574** records and **65** variables.

Categorical key variable(s): **REGION URBRUR GENDER MARITAL AGEYRS ATSCHOOL EDUCY INDUSTRY1**  
Sampling weight: **WGTPOP**  
Hierarchical identifier: **IDH**  
Deleted variable(s): **DIST**

### Computation time

The current computation time was ~ **11.13 seconds** .

### Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level
REGION	6 (6)	1762.333 (1762.333)	1348 (1348)
URBRUR	2 (2)	5287.000 (5287.000)	3486 (3486)
GENDER	2 (2)	5287.000 (5287.000)	5126 (5126)
MARITAL	7 (7)	1240.167 (1240.167)	295 (295)
AGEYRS	96 (96)	109.211 (109.211)	1 (1)



# Select a method

**sdcmicro GUI**    About/Help    Microdata    Anonymize    Risk/Utility    Export Data    Reproducibility    Undo

**View/Analyze existing sdcProblem**

- Show summary
- Explore variables
- Add linked variables
- Create new IDs

**Anonymize categorical variables**

- Recoding**
- k-Anonymity
- PRAM (simple)
- PRAM (expert)
- Suppress values with high risks

**Anonymize numerical variables**

## Recode categorical key variables

To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. You need to select a categorical key variable and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new category can be assigned.

Note: If you only select only one level, you can rename the selected value.

Choose factor variable  
MARITAL

Select levels to recode/combine

- married monogamous (2191 obs)
- married polygamous (440 obs)
- never married (3829 obs)
- Common law, union coutumiere, union libre, living together (295 obs)

Specify new label for recoded values  
married monogamous\_ma

Add missing values to new factor level?  
 no     yes

**Recode key variable**

### Variable selection

Variable name	Type	Additional suppression by local suppression algorithm
REGION	cat. key variable	0
URBRUR	cat. key variable	0
GENDER	cat. key variable	0
MARITAL	cat. key variable	0
AGEYRS	cat. key variable	0

# Display Risk and utility measures

sdcMicro GUI    About/Help    Microdata    Anonymize    **Risk/Utility**    Export Data    Reproducibility    Undo

**Risk measures**

Information of risk

Suda2 risk measure

I-Diversity risk measure

**Visualizations**

Barplot/Mosaicplot

Tabulations

Information loss

Obs. violating k-anon

**Numerical risk measures**

Compare summary statistics

Disclosure risk

Information loss

## Risk measures

The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

Risk measures     Risky observations     Plot of risk

### Risk measures

0 observations have a higher risk than the risk in the main part of the data, as compared to 0 observations in the original data ⓘ

Based on the individual re-identification risk, we expect **2.59 ( 0.02% )** re-identifications in the anonymized data set. In the original dataset we expected **2.59 ( 0.02% )** re-identifications.

### Hierarchical risk

The dataset has a hierarchical structure, which increases the risk of re-identification. ⓘ

If we take the hierarchical structure into account, the individual re-identification risk is the risk that at least one of the members is re-identified. If the hierarchal structure in the data is taken into account, we expect **12.64 ( 0.12% )** re-identifications in the anonymized data set. In the original dataset we expected **12.64 ( 0.12% )** re-identifications.

### Variable selection

Variable name	Type	Additional suppression by local suppression algorithm
REGION	cat. key variable	0
URBRUR	cat. key variable	0
GENDER	cat. key variable	0
MARITAL	cat. key variable	0
AGEYRS	cat. key variable	0
ATSCHOOL	cat. key variable	0

# Export the anonymized data

sdcmicro GUI

About/Help Microdata Anonymize Risk/Utility **Export Data** Reproducibility Undo

What do you want to export?

- Anonymized Data**
- Anonymization Report
- Change Stata Labels

Region 4	Urban	425.038421630859	2975.26904296875	1	7	7	Female	Grandparent
Region 3	Rural	331.750061035156	1327.00024414062	2	1	4	Male	Head
Region 3	Rural	331.750061035156	1327.00024414062	2	2	4	Female	Spouse
Region 3	Rural	331.750061035156	1327.00024414062	2	3	4	Female	Child

Showing 1 to 10 of 10,574 entries

Previous **1** 2 3 4 5 ... 1058 Next

Select file format for export

R-dataset (.RData)  SPSS-file (.sav)  CSV-file (.csv)  STATA-file (.dta)  SAS-file (.sas7bdat)

Randomize order of records ⓘ

No randomization  Randomize by hierarchical identifier  Randomize by hierarchical identifier and within hierarchical units

**Save dataset**



# Reproducibility

sdcMicro GUI

About/Help Microdata Anonymize Risk/Utility Export Data **Reproducibility** Undo

What do you want to do?

- [View the current script](#)
- Import a previously saved problem
- Export/Save the current sdcProblem

## 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.

[Save Script to File](#)

```
require(sdcMicro)
inputdata <- readMicrodata(path="/private/var/folders/bz/mrdgnf8s463_p4__7hxxy0400000gn/T/RtmpiM57QG/883256bd124a19470ac89307/case_1_data_lab.dta", type="stata", convertCharToFac=TRUE, drop_all_missings=TRUE)
inputdataB <- inputdata

## Convert a numeric variable to factor (each distinct value becomes a factor level)
inputdata <- varToFactor(obj=inputdata, var=c("AGEYRS"))
## Set up sdcMicro object
sdcObj <- createSdcObj(dat=inputdata,
  keyVars=c("REGION","URBRUR","GENDER","MARITAL","AGEYRS","ATSCHOOL","EDUCY","INDUSTRY1"),
  numVars=NULL,
  weightVar=c("WGTPOP"),
  hhId=c("IDH"),
  strataVar=NULL,
  pramVars=NULL,
  excludeVars=c("DIST"),
  seed=0,
  randomizeRecords=FALSE,
  alpha=c(1))

## Store name of uploaded file
```



# Undo

[sdcMicro GUI](#)

[About/Help](#)

[Microdata](#)

[Anonymize](#)

[Risk/Utility](#)

[Export Data](#)

[Reproducibility](#)

[Undo](#)

## Undo last step

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

Recoded "MARITAL": "married monogamous", "married polygamous" to "married monogamous\_married polygamous"

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!

# Summary

- The use of R for many users, especially in developing country statistical agencies, is relatively new.
- Many staff in agencies would like to apply SDC methods but do not have the necessary R skills to use sdcMicro.
- These users benefit from a friendly GUI for sdcMicro.
- Allows users to immediately apply methods without knowing R
- Has proved a valuable training tool for the World Bank

# Resources and Contacts

Resources:

- [Statistical Disclosure Control for Microdata: A Practice Guide](#)
- [sdApp manual](#)
- Alexander Kowarik, Matthias Templ and Bernhard Meindl, R-Package “sdcMicro” (2017), URL: <https://cran.r-project.org/package=sdcMicro>
- [Matthew Welch](#), Senior Statistician, World Bank

Thank you