Exercises TRS with R

ESTP Course on Statistical disclosure control (SDC) methods and tools for census 2021

Johannes Gussenbauer

Eurostat

Goal of exercises

- Load necessary packages and test data
- Transform variables to make test data
- Apply targeted record swapping on test data
- Use different set of parameters and compare resulting tables

The dataset we will be using in the exercises is `test_data_100k.csv.gz`. It is available for download in the "Files" tab of the "General" channel in the folder "R Testdata".

In addition there is a meta data file "Metadata_test_data.xlsx" with variable descriptions. Start RStudio (or R) on your PC or Laptop and create a new .R-File:

```
- File -> new File -> R Script
```

Write the code needed to solve the exercises in this file. Save this file in the same directory as `test_data_100k.csv.gz`

```
- File -> Save as -> Path to test_data_100k.csv.gz
```

Exercise 1 (2)

1. Load packages and data

```
# load packages
library(data.table)
library(recordSwapping)


# set the working directory
# setwd("path_to_test_data_100k.csv.gz")


# read data
dat <- fread("test_data_100k.csv.gz")
```

Exercise 1 (3)

1.a Make yourself familiar with the data at hand

- Print the data to the console
- Use `View()` for a direct look at the data
- Look at the meta data file "Metadata_test_data.xlsx"
- Count number of persons and households and number ~ nrow(data), uniqueN(data$HID)
- Check which columns are not of type integer

1.b Transform each column to integer and set an upper bound of 5 for variable Size

- The grid variable L001000 has the structure `1kmN`*Y-Coord*`E`*X-Coord*

1.c Save data using fwrite() for later use with muArgus.

Exercise 2

2.a Use the function recordSwap() to apply TRS with the following parameter

```
hierarchy <- c("NUTS1","NUTS2")
hid <- "HID"
risk_variables <- c("COC.M","POB.M")
k_anonymity <- 3
swaprate <- 0.05
similar <- "Size"
seed <- 2021
```

2.b Calculate how many households have been swapped

Exercise 2

2.c Check the swapped variables `NUTS2`, `NUTS3` and `LAU1` for consistency

2.d Use the parameter carry_along and repeat exercis 2.a to swap all geographic variables `NUTS3`, `LAU2`, X and Y grid coordinates.

- Check again if geographic variables are consistent

Exercise 3

3.a Construct table 8.1 using the original and swapped data (from 2.d) with the `data.table`-syntax, e.g `dat[,.N,by=c(...)]`

- `8.1 ~ NUTS3 x SEX x AGE.M x POB.H`

3.b Calculate information loss using AD, RAD and HD

- `Estimate summary statistics for each NUTS3 region`
- `Average over estimates for each NUTS3 region`

Exercise 3 (2)

3.c Use the function cellKey::ck_cnt_measures() to get information loss and look at the results

3.d Repeat the exercise and set the swaprate to 10% and 2.5%

3.f Compare the results for the different calls for `recordSwap()`

Exercise 4

4.a Construct tables 26.2 and 26.3 using the original and swapped data (from 2.d)

- `26.2 ~ NUTS2 x SEX x AGE.M x HST x POB.L`
- `26.3 ~ NUTS1 x SEX x AGE.M x HST x COC.L x POB.L`

Exercise 4 (2)

4.b Calculate information loss using AD, RAD and HD

- Estimate summary statistics for aggregates in each NUTS2/NUTS1 region
- For RAD estimate summary statistics for cross tables defined by NUTS2/NUTS1 x HST x POB.L
- Why is the information loss „Inf"?

4.c Similarly to exercise 3.c use cellKey::ck_cnt_measures().

Exercise 4 (3)

4.d Apply record swapping using the parameter set from exercise 2.d and set parameter `similar` to `c("Size","HST")`. Compare information loss with previous results.