

Statistiker:innen Treffen 2026:

Workshop on data anonymization: Exercise

Reto Bürigin, bure@zhaw.ch

Institut for Data Science, ZHAW

2026-09-08

We are considering a modified and reduced version of the survey data from the Age Monitor by Pro Senectute Switzerland from the year 2022. The data is stored in the file *ProSenectutePerturbed.RData*, see workshop materials.

```
load(file = "ProSenectutePerturbed.RData")
```

The data `sas.22_1_reduced` include 4'457 data records on 8 variables:

	Bezeichnung	Beschrieb
1	<code>Respondent_ID</code>	Person ID
2	<code>canton_22</code>	Canton
3	<code>pi_sex_22</code>	Gender
4	<code>pi_ageinterview_22</code>	Age in years
5	<code>ac_physact_22_freq</code>	Frequency of physical activities
6	<code>ac_cult_22_freq</code>	Frequency of cultural activities
7	<code>ac_socact_22_freq</code>	Frequency of social activities
8	<code>ac_concentrate_22_freq</code>	Frequency of activities that require strong concentration

An official version of the data is available from SWISSUbase.

```
head(sas.22_1_reduced, 3)
##   Respondent_ID canton_22 pi_sex_22 pi_ageinterview_22      ac_physact_22_freq
## 1          10000         LU  männlich                69 Jeden Tag oder fast jeden Tag
## 2          10002         TG  weiblich                77 Jeden Tag oder fast jeden Tag
## 3          10004         GE  weiblich                81 Jeden Tag oder fast jeden Tag
##               ac_cult_22_freq          ac_socact_22_freq      ac_concentrate_22_freq
## 1 Mindestens einmal pro Monat Mindestens einmal pro Monat Mindestens einmal pro Woche
## 2 Mindestens einmal pro Jahr Jeden Tag oder fast jeden Tag Jeden Tag oder fast jeden Tag
## 3               Nie Mindestens einmal pro Monat Mindestens einmal pro Woche
```

Goal: The goal of this task is to anonymize the Pro Senectute data with respect to the attribute disclosure scenario by coarsening variables. The data should be adjusted so that it satisfies the k -anonymity criterion with $k = 5$.

```
install.packages("sdcMicro")  
library(sdcMicro)
```

sdcMicro includes a web application, see

<https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdcMicro.html>

```
sdcApp()
```

Tabs of web application:

- **About/Help:** Specify the directory where the results should be stored
- **Microdata:**
 1. *Select data source > R-Dataset (.rdata)* $\Rightarrow$ Load the file *ProSenectutePerturbed.RData*
 2. *Explore variables* $\Rightarrow$ Inspect the distributions of the variables
 3. *Set specific values to NA* $\Rightarrow$ Recode the value “andere” to NA
Why might this be required?
 4. *Convert variables to numeric* $\Rightarrow$ Convert age into 5-year classes between 50 and 105
Because with did not discuss numeric key variables.
- **Anonymize:** ...

Tabs of web application:

- **About/Help:** ...
- **Microdata:** ...
- **Anonymize:**
 1. Define the anonymization task: Remove `Respondent_ID`, define `canton_22` and `pi_sex_22` and `pi_ageinterview_22` as categorical key variables, then click *Setup SDC problem*
 2. A summary page appears. The data currently do **not** satisfy 5-anonymity
 3. *Anonymize categorical variables > Recoding* Recode one of the key variables to increase anonymity
 4. *View/Analyze existing sdcProblem > Show Summary* Check whether 5-anonymity is now satisfied
 5. Repeat steps 3 and 4 until the data satisfy the 5-anonymity

Solution: See workshop materials, *docAnonymisierung_simple.pdf* (German)