Train the Trainer: Anonymisation for data sharing in practice - Additional Materials

Additional Resources

https://ukanon.net/framework/ - Website of the UK Anonymisation Network, includes the Anonymisation Decision Making Framework (ADF) and many additional tools and companion documents (e.g. Data Situation Evaluation and Data Features template).

https://dmeg.cessda.eu/Data-Management-Expert-Guide - CESSDA Data Management Expert Guide which contains numerous helpful guidelines on data management and a section dedicated to anonymisation (under chapter 5. Protect) where you can find a table categorising identifiers and easiest methods for anonymisation.

https://cpg.doc.ic.ac.uk/observatory/ - An interactive website made by the Computational Privacy Group at Imperial College London where you can take a short quiz to find out what makes you more vulnerable to re-identification and explore anonymity in 89 countries around the world. Useful for explaining the concept of singling out a research participant.

https://sdcappdocs.readthedocs.io/_/downloads/en/latest/pdf/ - Documentation for sdcMicro graphical interface by Thijs Benschop. Unfortunately, still missing descriptions for some procedures like PRAM, adding noise and some utility measures but nevertheless a great source for learning sdcMicro. Also includes two case studies with description of the anonymisation process.

https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf - Opinion 05/2014 on Anonymisation Techniques by the Data Protection Working Party which analyses the effectiveness and limits of existing anonymisation techniques against the EU legal background of data protection and provides recommendations to handle these techniques by taking account of the residual risk of identification inherent in each of them.

https://www.r-project.org/ - Official website of the R Project for Statistical Computing where you can find information about the R project and links for downloading R.

https://www.rstudio.com/ - Official website of RStudio, an IDE for R. Site includes download links and support documentation.

https://cran.r-project.org/web/packages/sdcMicro/index.html - The Comprehensive R Archive Network (CRAN) link for downloading sdcMicro.

https://www.hhs.gov/ohrp/sites/default/files/international-compilation-of-human-research-standards-2017.pdf - The International Compilation of Human Research Standards

enumerates over 1000 laws, regulations, and guidelines that govern human subjects research in 126 countries, as well as standards from a number of international and regional organisations. Although intended mainly for use by medical researchers, it includes links to Privacy and Data protections laws and regulation which can be useful for social scientists.

https://www.youtube.com/watch?v=KPFfxMba3Yg - CESSDA Train the Trainer workshop aimed at trainers and support staff looking to introduce semi-automated tools in data management training sessions/materials. Includes QAMyData (open source tool for numeric data health check, developed by the UK Data Service) presentation and a demo of sdcMicro.

https://zenodo.org/record/5566858 - Exercises and presentations for the CESSDA Train the Trainer workshop "Facilitating Onwards Sharing of Safe and Clean Microdata".

Exercises - sdcMicro

Exercise 1 - Installing sdcMicro

1. Open Rstudio (or alternative IDE)

2. Type in the Console window (by default left down) of your IDE ***install.packages("sdcMicro")*** and run the command by pressing Enter. If you want to also install packages needed by the sdcMicro type ***install.packages("sdcMicro", dependencies = TRUE)***

or

1. Open Rstudio (or alternative IDE)

2. (RStudio) Click on the "Packages" box in the Plots, files and help window (by default right down).

3. Choose Install from: Repository (CRAN), type "sdcMicro" under Packages and choose Install dependencies if you want to also install necessary packages.

Exercise 2 - Launching sdcMicro

1. Open RStudio (or alternative IDE)

2. Type in the Console window (by default left down in RStudio) of your IDE ***library (sdcMicro)*** to load the package.

3. Type in the Console window (by default left down in RStudio) of your IDE ***sdcApp()*** to launch the GUI app in your default browser.

Note: Installing the package sdcMicro is only required once, but loading the package is required every time a new R session is started. Good practice is to save a script containing these commands so you don't have to type them every time you want to start sdcMicro.

Exercise 3 - Loading data

1. Launch sdcMicro and sdcApp

2. Click on Microdata tab

3. Choose the right option on the left sidebar and select additional options for data import (if needed)

4. Click Browse... and select the file

5. Click Load data

6. After loading the dataset, it will be displayed (by default first 20 records)

Exercise 4 - Exploring variables

After loading the wanted dataset into sdcApp, the data is shown on the microdata tab. At the top of the data viewer, the number of observations and variables is shown as well as the number of variables that were dropped because of all missing values.

It is important to check whether the data was imported completely and correctly by browsing the dataset. If, for example, records are missing or labels are corrupted, then these issues need to be fixed outside of sdcApp and the data needs to be reimported.

Exercise 5 - Setting up your SDC problem

1. Navigate to the Anonymize menu and select key variables (tick the box).

2. Variables that are chosen as key variables in the SDC problem need to be marked as either categorical or continuous. If a variable is not a key variable, the default No should be selected. At least one variable needs to be selected as a categorical key variable in order to create an SDC problem.

3. Check if you need to select additional options (Weight, Hierarchical identifier, PRAM, Delete).

Exercise 6 - Recoding

A. *"Global" recoding* - combines several categories (levels) of a categorical variable or constructs intervals for continuous variables, thereby reducing the number of categories available in the data and likely the disclosure risk.

1. Navigate to the Anonymize tab and select Recoding from the left sidebar
2. Select the variables to be recoded
3. Select all the existing levels in the variable to be combined
4. Specify new label for recoded values (by default the label is created by concatenation of the labels of all combined values)
5. If the system missing value should be included in the newly created level, set the option Add missing values to new factor level to yes
6. Click Recode key variable

B. *Top/bottom coding* - similar to global recoding, but instead of recoding all values, only the top and/or bottom values of the distribution or categories are recoded. Top and bottom coding is especially useful if the bulk of the values lies in the centre of the distribution with the peripheral categories having only few observations (outliers).

1. Navigate to the Anonymize tab and select Top/bottom coding from the left sidebar
2. Select the variable to be recoded (only variables of type numeric can be top or bottom coded).
3. Select top or bottom coding.
4. Set threshold value
5. Set replacement value
6. Click Apply Top/Bottom-Coding

References

1. Benschop, T. (2021) sdcMicro GUI manual documentation.
2. Benschop, T., Machingauta, C., & Welch, M. (2019). Statistical disclosure control: A practice guide.
3. CESSDA Training Team (2017 - 2022). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Retrieved from https://dmeg.cessda.eu/
4. Elliot, M., Mackey, E., & O'Hara, K. (2020). The anonymisation decision-making framework 2nd Edition: European practitioners' guide.
5. Elliot, M., Mackey, E., O'Hara, K & Tudor, C. (2016). *The anonymisation decision-making framework*. UKAN Publications.
6. Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34*(2), 204-221.
7. EU Data Protection Working Party (2022, October 3) European Commission. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Materials created and compiled by Vedran Halamić, on behalf of Croatian Social Science Data Archive for the CESSDA "Train the Trainer Workshop: Anonymisation for data sharing in practice", October 2022