



Working with sdcMicro

From risky to shareable — anonymizing a real dataset, live

Prof. Dr. Matthias Templ

FHNW ■ SwissAnon ■ Lead
developer of sdcMicro

Data Protection Day 2026 ■
University of Zurich

Removing names is where it starts, not where it ends

- You stripped names, e-mails, birthdates. Done? **No** — that is where the real work begins.
- **Pseudonymization**: replace identifiers with artificial ones ("Hans Bauer" → "Participant1"). **Reversible** with the key.
- **Anonymization**: irreversible — data subjects can no longer be identified.
- Even with direct identifiers gone, **quasi-identifiers** (age, sex, location, ...) can re-identify records.
- Sweeney (2000): 87 % of Americans are unique on ZIP code, date of birth and sex.
- Under the GDPR, pseudonymized data is **still** personal data.

Every variable plays one role

- **Direct identifiers:** identify a person directly — names, passport numbers, addresses. *(Remove.)*
- **Quasi-identifiers:** age, sex, occupation, location, ... — do not identify alone, but **re-identify in combination** or via linkage to external data. *(Protect.)*
- **Confidential / sensitive variables:** health status, substance use, attitudes — what we must keep from leaking.
- **Non-confidential variables:** colour preference, reaction time — harmless on their own.

Measuring disclosure risk

k -Anonymity (Sweeney, 2002)

- Each combination of quasi-identifiers appears in at least k records.
- Nobody should be unique on the released quasi-identifiers ($k \geq 2$).
- Protects against **identity** disclosure (finding *who*).

ℓ -Diversity (Machanavajjhala et al., 2007)

- Each k -anonymous group must also be **diverse** in the sensitive variable.
- A group can satisfy k -anonymity yet reveal the same outcome for everyone in it.
- Protects against **attribute** disclosure (learning *what*).

Example: k -anonymity and ℓ -diversity

Fictitious infant dataset: a unique record ($k = 1$), a protected group ($k = 2, \ell = 2$), and a group with attribute disclosure ($k = 2, \ell = 1$).

Quasi-identifiers			Sensitive variable	k	ℓ
Age (months)	Race	Gender	Newborn health condition		
0–3	Asian	Female	Low birth weight	1	1
0–3	Black	Female	Colic	2	2
0–3	Black	Female	Low birth weight	2	2
4–6	White	Male	Lactose intolerance	2	1
4–6	White	Male	Lactose intolerance	2	1

The sdcMicro toolbox — three families we will use live

Today, on the Coleman data:

- **Recoding / generalization** — age bands, merge rare categories
- **Suppression** — replace the few remaining risky values with NA to reach k -anonymity
- **Perturbation** — PRAM: swap categories under an invariant transition matrix

Also in the box (mentioned, not shown):

- Rank swapping & added noise
- Synthetic data
- Record swapping, microaggregation, ...

Everything via the sdcMicro package (Templ et al., 2015) or its GUI sdcApp (Meindl & Templ, 2019).

Our case: the Coleman depression study

The data

- $n = 107$ participants, UK sample — real psychology data
- 4 quasi-identifiers: Age, Sex, Occupation, Ethnicity
- 47 sensitive Likert items $\rightarrow$ depression (PHQ-9), derailment, self-criticism, self-reassurance
- Direct identifiers already removed

Colman et al. (2024), PsychArchives; re-analysed in Müller, Templ & Strobl (2026).

Why this dataset

- The everyday case: **small- n human-subjects data**
- Small enough to read row by row
- Lets us watch k -anonymity **fail** — and fix it

Step 1 — write the disclosure scenario *first*

Our scenario (fixed before we touch a method)

Release mode **Public Use File** — openly downloadable, no data-use agreement.

Attacker knows Age, Sex, Occupation, Ethnicity of the target, and that they took part.

Harm identity disclosure $\Rightarrow$ the attacker reads the participant's **depression score**.

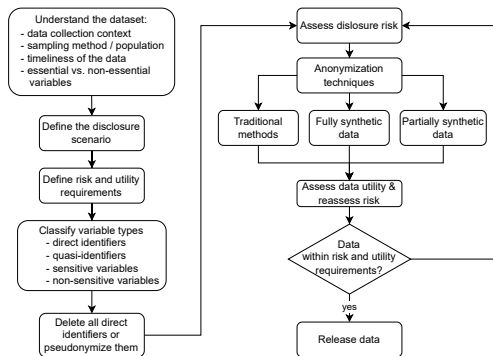
Utility the original regression must stay reproducible.

Without a written scenario you cannot decide *when to stop* — suppress one more cell? bands of 5 or 20 years?

56 % of these 107 people are unique on just those four variables.

Let us watch sdcMicro fix that — live.

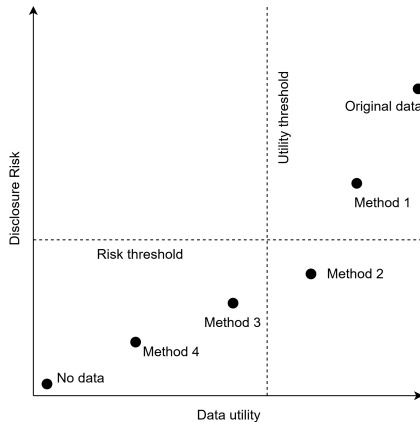
The workflow we run now — live in sdcApp



Risk → recode → suppress → perturb → check utility → document. (Switch to `sdcApp()` now.)

Risk vs. utility — there is no free lunch

- Every method trades information for protection.
- The goal is not zero risk, but an *acceptable balance*.
- The balance depends on the use case and the sensitivity of the data.
- Your written scenario is what makes "acceptable" concrete.



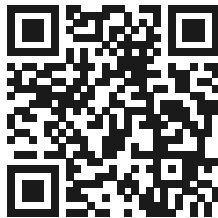
Take-home

- Removing names is **not** anonymization — quasi-identifiers remain.
- Write the **disclosure scenario** first; it tells you when to stop.
- **Recode** → **suppress** → **perturb**, then re-measure and document.
- sdcMicro gives you a *toolbox* and an *audit trail* — the report is the deliverable, not optional metadata.

Try it yourself

- `install.packages("sdcMicro")`
- `library(sdcMicro); sdcApp()` — launches the GUI
- Reports ship *with* the data: `report(sdc, internal = TRUE/FALSE)`

Go deeper: sdcMicro (Templ et al., 2015) ■ sdcApp GUI (Meindl & Templ, 2019) ■ Workshop 5 (data-sharing Q&A) ■ find me afterwards — FHNW / SwissAnon.



Slides, script & data

swissanon.com/dpd2026

References

- Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 1–52. <https://doi.org/10.1145/1217299.1217302>
- Meindl, B., & Templ, M. (2019). Feedback-based integration of the whole process of data anonymization in a graphical interface. *Algorithms*, 12(9), 191. <https://doi.org/10.3390/a12090191>
- Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Data Privacy Working Paper No. 3). Carnegie Mellon University. Pittsburgh, PA. Retrieved May 18, 2026, from <https://privacytools.seas.harvard.edu/files/privacytools/files/paper1.pdf>

References

- Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. <https://doi.org/10.1142/S0218488502001648>
- Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4), 1–36. <https://doi.org/10.18637/jss.v067.i04>

Thank you

Questions?



Backup — setup: assign the key variables

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

Anonymize
Select variables and set parameters to create the SOC problem.

Select variables

Variable name	Type	Key variables	Weight	Anonymized identifier	PSAM	Delete	Number of levels	Number of missing
Q0408	integer	☒ No ☐ Categorical	☐	☒	☐	☐	9999	0
h0408	integer	☐ No ☒ Categorical	☐	☐	☐	☐	9	0
o0408	factor	☒ No ☐ Categorical	☐	☐	☒	☐	9	0
r0408	integer	☒ No ☐ Categorical	☐	☐	☐	☐	14027	0
age	integer	☐ No ☒ Categorical	☐	☐	☐	☐	99	0
r0408	factor	☐ No ☒ Categorical	☐	☐	☐	☐	2	0
p0408	factor	☐ No ☒ Categorical	☐	☐	☐	☐	6	2719
p0229a	factor	☐ No ☒ Categorical	☐	☐	☐	☐	4	2719
py0408	numeric	☒ No ☐ Categorical	☐	☐	☐	☐	6454	2719
py0408	numeric	☒ No ☐ Categorical	☐	☐	☐	☐	3919	2719

Setup SOC problem

Set additional parameters

Parameter 'alpha'

Parameter 'seed'

Explore variables

Index [integer]

Number of levels including missing (Out): 10

Level	Frequency
1	1709
2	1624
3	1347
4	1598
5	1813
6	676
7	252
8	88
9	18
NA	0

Anonymize data → select Age, Sex, Occupation, Ethnicity as key variables.

Backup — initial risk: 56 % sample-unique

sdcMicro GUI

View/Help | Metadata | Anonymize | Risk/Utility | Export Data | Reproducibility | Undo

EXPOSED k -ANONYMITY

k -anonymity will be obtained by suppressing or other setting some values in the selected categorical key variables to **NA**.

By default, the key variables will be considered for suppression in the order of their number of distinct categories. A variable with many categories is less likely to have values suppressed than a variable with few categories. It is also possible to set the order by specifying an importance vector.

You may also decide to apply the procedure for all possible subsets of key variables. This is useful, if you have many key variables and can reduce computation time. You can set a different value for the parameter β for each subset of subsets.

Do you want to apply the method for each group defined by the selected variable?

Do you want to modify importance of key variables for suppression?

Tip: The total number of suppressions is likely to increase by specifying an importance vector. Specifying an importance vector can affect the computation time.

Select the importance for key variable "gender": 0 1 2 3 4 5 6 7 8 9 10

Select the importance for key variable "age": "NONE"

Select the importance for key variable "year": 0 1 2 3 4 5 6 7 8 9 10

Select the importance for key variable "year2": 0 1 2 3 4 5 6 7 8 9 10

Select the importance for key variable "year3": 0 1 2 3 4 5 6 7 8 9 10

Apply k -anonymity to subsets of key variables?

Set the k -anonymity parameter

Variable selection

Variable name	Type	Additional suppression by local suppression algorithm
year1	cat	NA
year2	cat	NA
year3	cat	NA
year4	cat	NA
year5	cat	NA
year6	cat	NA
year7	cat	NA
year8	cat	NA
year9	cat	NA
year10	cat	NA
year11	cat	NA
year12	cat	NA
year13	cat	NA
year14	cat	NA
year15	cat	NA
year16	cat	NA
year17	cat	NA
year18	cat	NA
year19	cat	NA
year20	cat	NA
year21	cat	NA
year22	cat	NA
year23	cat	NA
year24	cat	NA
year25	cat	NA
year26	cat	NA
year27	cat	NA
year28	cat	NA
year29	cat	NA
year30	cat	NA
year31	cat	NA
year32	cat	NA
year33	cat	NA
year34	cat	NA
year35	cat	NA
year36	cat	NA
year37	cat	NA
year38	cat	NA
year39	cat	NA
year40	cat	NA
year41	cat	NA
year42	cat	NA
year43	cat	NA
year44	cat	NA
year45	cat	NA
year46	cat	NA
year47	cat	NA
year48	cat	NA
year49	cat	NA
year50	cat	NA
year51	cat	NA
year52	cat	NA
year53	cat	NA
year54	cat	NA
year55	cat	NA
year56	cat	NA
year57	cat	NA
year58	cat	NA
year59	cat	NA
year60	cat	NA
year61	cat	NA
year62	cat	NA
year63	cat	NA
year64	cat	NA
year65	cat	NA
year66	cat	NA
year67	cat	NA
year68	cat	NA
year69	cat	NA
year70	cat	NA
year71	cat	NA
year72	cat	NA
year73	cat	NA
year74	cat	NA
year75	cat	NA
year76	cat	NA
year77	cat	NA
year78	cat	NA
year79	cat	NA
year80	cat	NA
year81	cat	NA
year82	cat	NA
year83	cat	NA
year84	cat	NA
year85	cat	NA
year86	cat	NA
year87	cat	NA
year88	cat	NA
year89	cat	NA
year90	cat	NA
year91	cat	NA
year92	cat	NA
year93	cat	NA
year94	cat	NA
year95	cat	NA
year96	cat	NA
year97	cat	NA
year98	cat	NA
year99	cat	NA
year100	cat	NA

Additional parameters

Parameter	Value
number of records	10000
alpha	0.1
random seed	0

k -ANONYMITY

k -anonymity	Modified data	Original data
2-anonymity	10000	10000
3-anonymity	10000	10000
4-anonymity	10000	10000
5-anonymity	10000	10000
6-anonymity	10000	10000
7-anonymity	10000	10000
8-anonymity	10000	10000
9-anonymity	10000	10000
10-anonymity	10000	10000

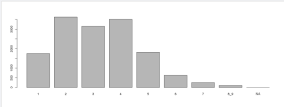
Risk in numerical key variables

Data	Minimum risk	Maximum risk
modified	0.00%	100.00%
original	0.00%	100.00%

Information loss

Measure	Modified data	Original data
Entropy	0.00	0.00
Difference in entropy	0.00	0.00

Risk $\rightarrow k$ -Anonymity. Red rows violate the chosen k .



Month	Percentage (%)
March 2020	45
April 2020	50
May 2020	55
June 2020	50
July 2020	52
August 2020	48
September 2020	50
October 2020	48
November 2020	50
December 2020	48
January 2021	50
February 2021	48
March 2021	45

Anonymize $\rightarrow$ *Local suppression*: set k and the importance vector, *Apply*.



Anonymize → *PRAM*: pick the variables, accept or edit the matrix, *Apply*.