

Illustration einer Anonymisierung von quantitativen Umfragedaten in R

Reto Bürgin¹, Michelle Haas¹, Tibor Pimentel¹, Hedi Strahm¹

^{aff-1}Zurich University of Applied Science (ZHAW)

3. September 2026

Dieses Dokument bietet Forschenden der ZHAW anhand einer Illustration mit realen Umfragedaten einen praxisbezogenen Zugang zur Thematik der Datenanonymisierung. Schritt für Schritt werden sensible Ausgangsdaten modifiziert, so dass diese auf Anfrage weitergegeben oder gar publiziert werden könnten. Für die Anonymisierung der Umfragedaten werden ausschliesslich einfache Kriterien wie k-Anonymity zur Beurteilung des Reidentifikationsrisikos und einfache Anonymisierungsmethoden wie das Weglassen oder Vergrößern von Variablen verwendet.

Inhaltsverzeichnis

1	Einleitung	3
2	Die Pro Senectute Daten	4
3	Variablentypen und Reidentifikationsszenarios	5
4	Messung des Reidentifikationsrisikos	7
5	Anonymisierungsmethoden	8
6	Anonymisierung der Pro Senectute Daten	8
6.1	About/Help	10
6.2	Microdata	10
6.3	Anonymize	13
6.3.1	Iteration 1	14
6.3.2	Iteration 2	16
6.3.3	Iteration 3	17
6.4	Export Data	18
6.5	Reproducibility	19
7	Data Utility und Informationsverlust	20
8	Diskussion und Fazit	21
	Finanzierung	22
	Dankessagungen	22
	Referenzen	22
	Liste mit mit Beispielen zu typischen Quasi-Identifiers	22
	Anhänge	23

1 Einleitung

Gemäss Schweizerischem Nationalfonds SNF und Europäischer Kommission wird erwartet, dass Forschungsdaten aus von ihnen geförderten Projekten öffentlich zugänglich gemacht werden, sofern dies rechtlich, ethisch und urheberrechtlich möglich ist.^{1 2}

Um die Anonymität von quantitativen Umfragedaten zu gewährleisten braucht es oft Anpassungen an den Daten. Dazu gehört beispielsweise das Weglassen von Informationen (bspw. Vornamen, Namen), die Reduktion des Informationsgehalts (bspw. Einkommensklassen statt Einkommen in Franken), das Löschen von einzelnen Werten (bspw. ein Monatseinkommen von über 500'000 Fr. durch einen fehlenden Wert ersetzen) oder das Aggregieren von Daten (bspw. Daten auf Gemeinde- statt Individualebene angeben). Solche Anonymisierungsarbeiten stellen für Forschende oft eine Herausforderung dar und bringen einen erheblichen Zeitaufwand mit sich. Dieses Dokument bietet Forschenden der ZHAW einen praxisbezogenen Zugang zur Anonymisierung von Daten. Die Dokumentinhalte sind insbesondere:

- Illustration einer Anonymisierung an realen, quantitativen Umfragedaten
- Verwendung einfacher Kriterien zur Beurteilung des Reidentifikationsrisikos
- Verwendung einfacher Anonymisierungsmethoden. Diese umfassen insbesondere:
 - Weglassen von Variablen
 - Vergrößern der Werte
 - Löschen von einzelnen Werten

Die Anonymisierung orientiert sich grob am Ablauf von Abbildung 1, wobei die einzelnen Punkte in den nachfolgenden Kapiteln genauer beschrieben werden:

¹Siehe <https://www.snf.ch/de/dMILj9t4LNk8NwyR/thema/open-research-data>

²Siehe bspw. <https://library.ethz.ch/forschen-und-publizieren/datenmanagement-und-policies/forschungsfoerderung/europaeische-kommission.html>

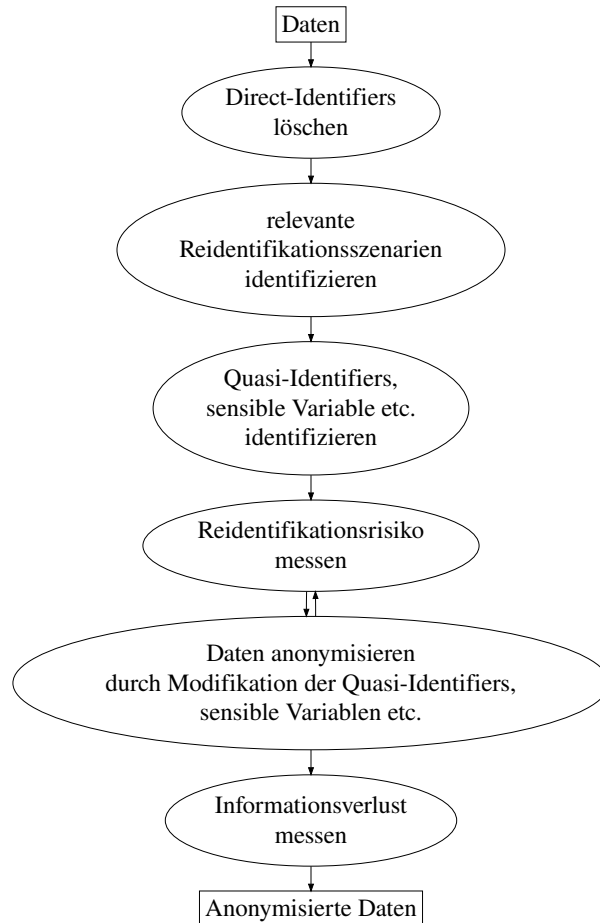


Abbildung 1: Allgemeiner Ablauf einer Anonymisierung.

Aus Sicht des Datenschutzes ist es nicht von Belang, ob die Daten mit einfachen oder ausgefeilten Methoden anonymisiert wurden. In jedem Fall können anonymisierte Daten generiert werden, welche Vorgaben zum Reidentifikationsrisiko wie k -Anonymity einhalten. Die einfachen Methoden haben den Vorteil, dass sie für Nichtfachkundige niederschwelliger und die Datenanpassungen sehr gut kontrollierbar sind.³ Dafür können ausgefeilte Methoden zeitaufwändige und komplexe Anonymisierungsschritte, wie beispielsweise das simultane Vergrößern von Werten bei mehreren Variablen, sehr viel effizienter erledigen, so dass für Geübte weniger Aufwand anfällt und der Informationsverlust geringer bzw. der Nutzwert der Daten grösser ist. Entsprechend empfehlen wir der Leserschaft den Deep Dive in die Literatur zum Thema, bspw. Templ (2017). In jedem Fall gibt es jedoch, zumindest unseren Kenntnissen nach, keine allgemeine Knopfdrucklösung für das Anonymisieren von Daten.

2 Die Pro Senectute Daten

Wir betrachten die Umfrage zum Altersmonitor von Pro Senectute Schweiz aus dem Jahr 2022 (Gabriel und Kubat 2022). Es handelt sich um eine repräsentative Umfrage zu verschiedenen

³Bspw. bei einer Vergrößerung des Alters in Jahren in Altersklassen können die Grenzen selbst bestimmt werden. Bei ausgefeilten Methoden ist die Vergrößerung meist automatisiert und kann “unintuitive” Klassen, wie “27 bis 32” oder ähnlich, generieren.

altersrelevanten Themen, die von der Pro Senectute Schweiz in Zusammenarbeit mit der ZHAW und der Universität Genf erstellt wurde. Die aus der Umfrage resultierenden Individualdaten umfassen viele sensible Informationen, beispielsweise zur physischen Gesundheit oder zu den Finanzen.

Die verwendeten Daten umfassen Informationen zu 4 457 Personen über 50 und umfassen 120 Variablen. Die Variablen können in die folgenden Blocks unterteilt werden:

Kürzel	Beschrieb	Anz. Variablen
ENV	Informationen zur räumlichen Umgebung	9
PI	Personeninformationen	16
SN	Soziales Netz	2
LS	Lebensqualität	6
CO	Kognition	1
PH	Physische Gesundheit	5
CS	Care & Support	7
MH	Mental Health	5
AC	Activities und Freizeitgestaltung	4
NT	Neue Technologien	6
SE	Sozioökonomische Position	11
FI	Finanzen	19
PS	Fragen zu Pro Senectute	21

Die Variablennamen verwenden jeweils das Kürzel des zugehörigen Variablenblocks als Prefix, bspw. das Geschlecht des Variablenblocks Personeninformationen (PI) trägt den Variablennamen `pi_sex_22`.

Eine anonymisierte Fassung dieser Daten wurde bereits vor dem Verfassen dieses Berichts erstellt. Bei dieser wurden sensitive Variablen prinzipiell weggelassen. Sie sind voraussichtlich ab Anfang 2025 via [SWISSUbase](#) verfügbar.

3 Variablentypen und Reidentifikationsszenarios

Für die Anonymisierung der Pro Senectute Daten sind folgende zwei Variablentypen relevant:

1. **Direct-Identifiers:** Variablen, die direkt und eindeutig die Identität der befragten Person offenbaren. Beispiele sind Namen, Reisepassnummern, Sozialversicherungsnummern und Adressen. Direct Identifiers werden bei der Anonymisierung von den Daten entfernt.
2. **Quasi-Identifiers:** Zusatzinformationen über die Umfrageteilnehmenden, die Angreifen potenziell vorliegen und in Kombination mit anderen Quasi-Identifiers zur Reidentifizierung genutzt werden können. Beispiele für Quasi-Identifiers sind ethnische Herkunft, Geburtsdatum, Geschlecht und Postleitzahlen. Quasi-Identifiers werden bei der Anonymisierung je nach dem belassen, modifiziert oder sogar entfernt. Eine Liste mit Beispielen ist im Anhang des Dokuments zu finden.

Weitere Variablentypen im Kontext von Anonymisierungen sind **sensible Variablen** (vertrauliche Informationen) oder **Design Variablen** (Informationen über die Datenerhebung). Im

Folgenden gehen wir davon aus, dass die Direct-Identifiers der vorliegenden Daten identifiziert und entfernt wurden.

Für das Anonymisieren von Daten ist es zunächst grundlegend, mögliche Szenarien für das zu unterbindende Reidentifizieren einzelner Personen festzulegen. Für die Pro Senectute Daten gehen wir vom sogenannten **Identity-Disclosure-Szenario** aus. In diesem verfügen die Angreifenden über Fremddaten mit Direct-Identifiers und eine Reihe an Quasi-Identifiers, die in den Daten vorhanden sind. Durch eine Datenverknüpfung mit diesen Fremddaten können die Umfrageteilnehmenden reidentifiziert werden. Abbildung 2 illustriert eine solche Datenverknüpfung grafisch.

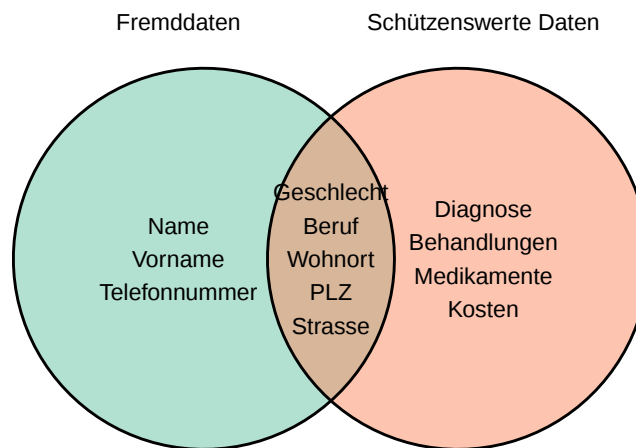


Abbildung 2: Illustration Datenverknüpfung zur Reidentifikation

Ein weiteres bedeutsames Szenario ist das sogenannte **Attribute-Disclosure-Szenario**. Bei diesem wird tiefe Variation in den Daten ausgenutzt, um für bestimmte Personen die Ausprägung bei sensiblen Variablen zu ermitteln. Mehr zu Reidentifikationszenarios findet sich in Templ (2017), Kap. 2.2.

Um die Daten hinsichtlich des Identity-Disclosure-Szenarios zu anonymisieren müssen zunächst die Quasi-Identifiers identifiziert werden. Direct Identifiers haben die Pro Senectute Daten ohnehin nicht, da diese anonym erhoben wurden. Als Quasi-Identifiers wurden insgesamt 16 Variablen eingestuft:

- Geografische Angaben zur Gemeinde (insgesamt 10 Variablen): Nummer des Bundesamts für Statistik (BfS) zur Wohngemeinde, Postleitzahl, Zugehöriger Kanton, Grossregion, Agglomerationsgrössenklasse, statistische Stadt (ja/nein)⁴, zwei Gemeindetypologien des BfS (9 und 25 Typen), Sprachgebiet, Urbanisierungsgrad, Berggebiet (ja/nein), Gemeindegrosse
- Interviewsprache
- Alter
- Geschlecht
- Schweizer Nationalität (ja/nein)

⁴Siehe Definition auf <https://www.bfs.admin.ch/asset/de/5931172>,

- Zivilstand
- Anzahl Kinder

Die obige Auswahl von Quasi-Identifiern ist mehrdeutig und hat Diskussionspotenzial. Beispielsweise könnte die Anzahl Kinder weggelassen oder Variablen zur sozioökonomischen Position, wie der Bildungsabschluss oder die Berufsbezeichnung hinzugefügt werden. Dabei kommt es stark auf den Grad der Veröffentlichung an: Werden die Daten als Open Data zur Verfügung gestellt, dann werden zur verbesserten Gewährleistung der Anonymität weniger Quasi-Identifier berücksichtigt als wenn lediglich geplant ist, die Daten eingeschränkt und mit Datenschutzvertrag mit Forschenden zu teilen.

4 Messung des Reidentifikationsrisikos

k-Anonymity ist ein weit verbreitetes Risikomass für das Identity-Disclosure-Szenario. Es beruht auf dem Grundsatz, dass in einem sicheren Datensatz die Zahl der Personen, welche dieselbe Kombination von Ausprägungen bei den Quasi-Identifiern (sogenannte Keys) teilen, gleich oder höher sein sollte als ein bestimmter Schwellenwert k . Je höher das k , das von allen Keys eingehalten wird, desto tiefer das Reidentifikationsrisiko bei den Daten. Zur Erfassung des Reidentifikationsrisikos wird berechnet, für welche (relative oder absolute) Anzahl an Personen eine bestimmte k -Vorgabe verletzt wird. Bei der Anonymisierung werden die Quasi-Identifier solange angepasst oder entfernt, bis ein bestimmtes k für alle oder fast alle Personen eingehalten wird.

Die nachfolgende Tabelle 1 zeigt fiktive Daten mit 5 befragten Personen. Wenn man beide Variablen Geschlecht und Altersklasse als Quasi-Identifiers festlegt, ergeben sich 2 Keys, nämlich Männer in den 30ern und Frauen in den 20ern. Da es $f_j = 3$ Männer in den 30ern und $f_j = 2$ Frauen in den 20ern gibt, lässt sich schliessen, dass lediglich $k=1$ und $k=2$ eingehalten wird. $k=3$ wird vom Key “Frauen in den 20ern” verletzt, und $k=4$ und höher wird von allen Keys verletzt.

Tabelle 1: Beispieldaten zu Patienten zur Illustration von k-Anonymity.

	Geschlecht	Altersgruppe	f_j
1	männlich	30er	3
2	männlich	30er	3
3	männlich	30er	3
4	weiblich	20er	2
5	weiblich	20er	2

Für k werden oft 3, 5 oder 10 eingesetzt, wobei auch höhere Werte möglich sind. Mit wachsenden k sinkt das Reidentifikationsrisiko, aber auch der Informationsgehalt und damit der Nutzwert der Daten. Somit ist die Wahl von k immer ein Kompromiss zwischen Sicherheit und Informationsgehalt. Für die Anonymisierung der Pro Senectute wird $k=5$ angestrebt, d.h. alle Keys sollen mindestens 5 Mal vorkommen.

Bemerkungen:

- k-Anonymity bezieht sich auf kategorielle Quasi-Identifiers, währenddem es für numerische Quasi-Identifiers alternative Risikomasse wie Record Linkage oder Interval Measure gibt, siehe bspw. Templ (2017), Kap. 3.11. Der Einfachheit halber werden bei den nachfolgenden Analysen alle numerischen Quasi-Identifiers zu kategoriellen Variablen umgewandelt, so dass die Betrachtung von k-Anonymity ausreicht.
- Für jede Art von Reidentifikationszenario gibt es spezifische Risikomasse. Beispielsweise wird beim Attribute-Disclosure-Szenario die sogenannte l-Diversity eingesetzt, und nicht k-Anonymity. l-Diversity misst, wie stark die Ausprägungen von sensitiven Variablen innerhalb von Keys variieren. Für eine allgemeine Einführung und Übersicht zur Risikomesung verweisen wir auf Templ (2017), Kap. 3.

5 Anonymisierungsmethoden

Nachfolgend werden zur Anonymisierung zwei einfache Techniken verwendet:

- **Weglassen von Variablen:** Die einfachste Methode zur Reduktion von Informationen, führt aber oft zu einem enormen Informationsverlust. Die Methode wird von daher vor allem in Situationen angewendet, wo gleichzeitig das Reidentifikationsrisiko stark gesenkt werden kann und die Variablen nur einen begrenzten Nutzwert haben.
- **Vergrößern von Variablen:** Modifikation der Variablen zur Schmälerung des Informationsgehalts, beispielsweise numerische Variablen in ordinal kategorielle Variable umwandeln (Alter in Jahren zu Altersklassen). Bei kategoriellen Variablen werden gezielt Kategorien zusammengelegt (bspw. Tessin und Romandie in Lateinische Schweiz).

Eine weitere einfache Anonymisierungsmethode ist die Post Randomization Methode. Dabei werden Werte zufällig vertauscht, z.B. werden in Einzelfällen Männer zu Frauen, und umgekehrt. Bei der sogenannten Local Suppression Methode werden hingegen Werte einfach gelöscht bzw. durch ein Missing Value ersetzt. Eine allgemeine Einführung und Übersicht zum Thema Anonymisierungsmethoden findet sich bspw. in Templ (2017), Kap. 4.

6 Anonymisierung der Pro Senectute Daten

Zur Vereinfachung der Illustration werden die Pro Senectute Daten auf die folgenden Variablen reduziert:

- ID (Variable `Respondent_ID`)
- Kanton (Variable `canton_22`)
- Alter (Variable `pi_ageinterview_22`)
- Geschlecht (Variable `pi_sex_22`)
- Alle Variablen des Variablenblocks Activities und Freizeitgestaltung (AC)

Als Quasi-Identifiers werden der Kanton, das Alter und das Geschlecht eingestuft. Die Daten sind der Leserschaft verfügbar als *ProSenectutePerturbed.RData*, z.B. um die folgende Anonymisierung zu reproduzieren. Bei diesen reduzierten Daten sind die Variablen des Variablenblocks AC perturbiert. Dies hat auf die Anonymisierung keine Auswirkungen, da die Variablen des Variablenblocks AC nicht als Quasi-Identifiers eingestuft werden.

Die Anonymisierung der Pro Senectute Daten wird mit der Statistiksoftware [R](#) und mithilfe des R-Pakets [sdcMicro](#) durchgeführt. sdcMicro (Abkürzung für: “Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation”) verfügt über viele Funktionen zur Messung des Reidentifikationsrisikos und zur Anonymisierung von Daten. Die Anwendung der Funktionen ist neben der Hilfe im Buch Templ (2017) dokumentiert. Das R-Paket sdcMicro umfasst für Anwender ohne R-Kenntnisse eine Webanwendung, siehe <https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdcMicro.html>, die auch hier verwendet wird. Die Webanwendung lässt sich nach Installation und Laden des R-Pakets via

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

mit dem R-Befehl `sdcApp()` aufrufen:

```
sdcApp()
```

Abbildung 3 zeigt die Startseite der Webanwendung. Auf dieser kann unter anderem eingestellt werden, wo Outputs gespeichert werden sollen. Ausserdem kann hier die Anwendung gestoppt werden.

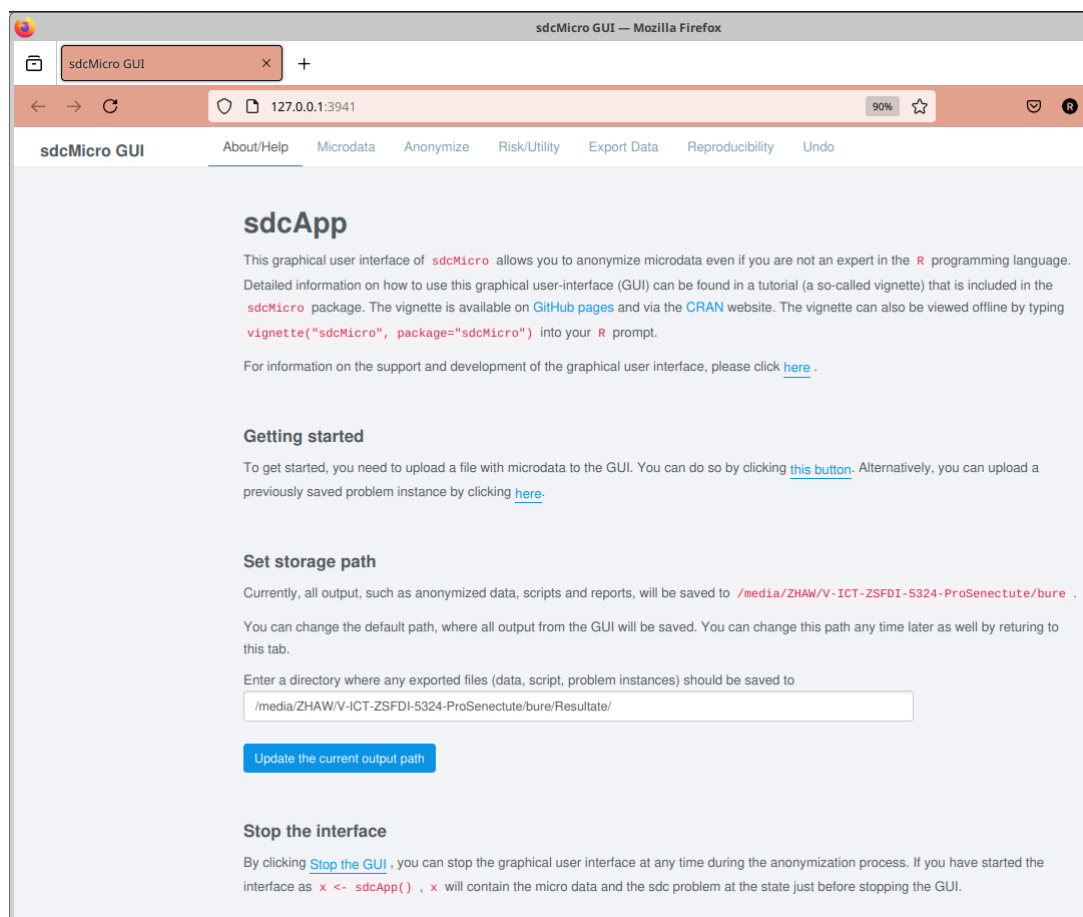


Abbildung 3: Startseite der Webanwendung `sdcApp()`.

Im Folgenden beschränken wir uns auf das Anonymisieren mit den 5 Reitern About/Help, Microdata, Anonymize, Export Data und Reproducibility.

6.1 About/Help

Zunächst legen wir den Pfad für die Ablage der Resultate fest, siehe Abbildung 4.

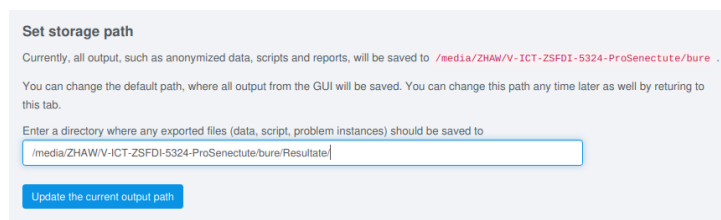


Abbildung 4: Pfad für Resultate festlegen.

Ausserdem beenden wir nach der Anonymisierung die Webanwendung mittels dem Link “Stop the GUI”.

6.2 Microdata

Nun werden die Daten eingelesen. In unserem Fall wurden die Daten vorgängig in R aufbereitet und liegen im R-eigenen RDATA-Format vor, was links im Menü spezifiziert werden muss. Abbildung 5 zeigt, wie die Daten eingelesen werden.

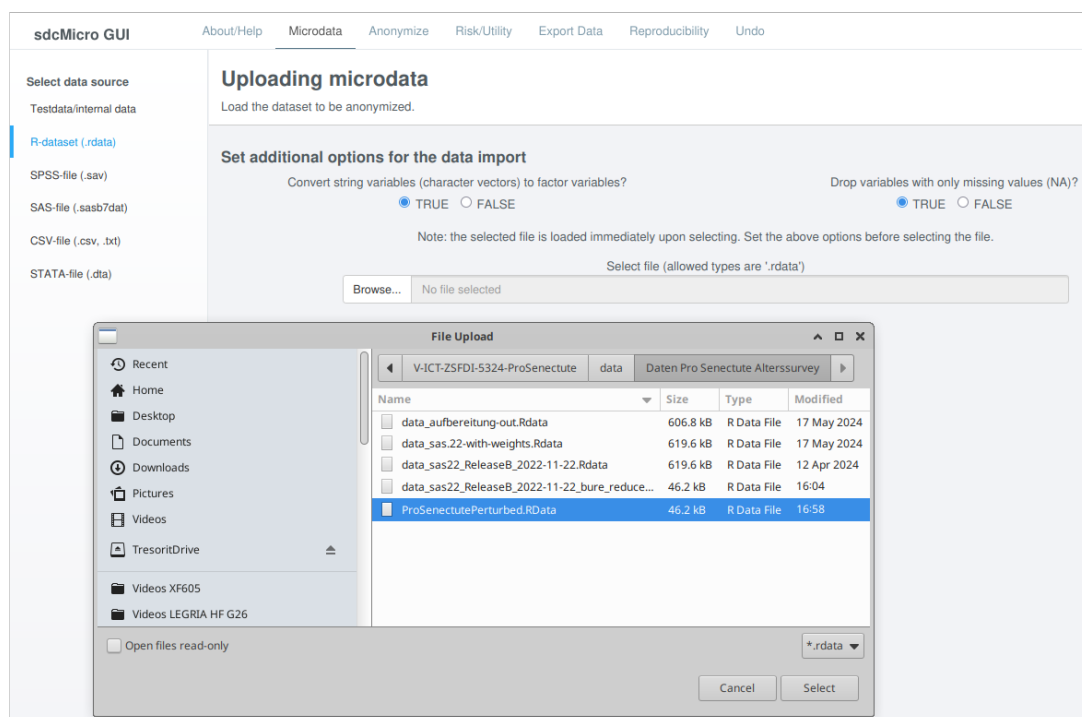


Abbildung 5: Daten einlesen.

Nach dem Einlesen der Daten öffnet sich innerhalb des Reiters eine neue Oberfläche, welche die explorative Analyse der Daten sowie erste Anpassungen an den Daten erlaubt, siehe Abbildung 6.

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

Loaded microdata

The loaded dataset is **ProSenectutePerturbed.RData** and consists of **4457** observations and **8** variables. No variables were dropped because of all missing values.

Show **20** entries Search:

Respondent_ID	canton_22	pi_sex_22	pi_ageinterview_22	ac_physact_22_freq	ac_cult_22_freq	ac_socact_22_freq	ac_concentrate_22
10000	LU	männlich	69	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Monat	Jeden Tag oder fast jeden Tag	Jeden Tag oder jeden Tag
10002	TG	weiblich	77	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Jahr		Jeden Tag oder jeden Tag
10004	GE	weiblich	81	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Jahr	Mindestens einmal pro Monat	Mindestens eine Woche
10005	NW	weiblich	78	Nie	Mindestens einmal pro Jahr	Mindestens einmal pro Woche	Mindestens eine Woche

Showing 1 to 20 of 4,457 entries Previous 1 2 3 4 5 ... 223 Next

Abbildung 6: Erste Analysen und Anpassungen an den Daten.

Da die eingelesenen Daten bereinigt sind, werden lediglich zwei Datenanpassungen gemacht. Erstens wird unter “Explore variables” ersichtlich, dass bei der Geschlechtsvariable **pi_sex_22** die Kategorie “andere” nur einmal vorkommt, siehe Abbildung 7.

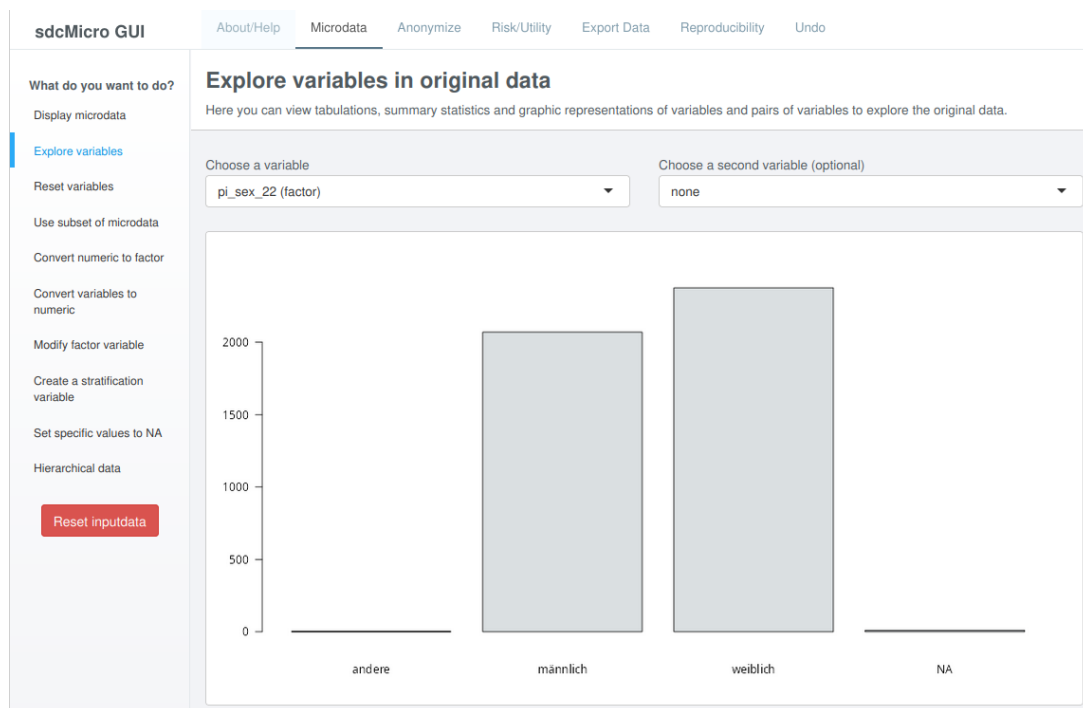


Abbildung 7: Verteilung der Geschlechtsvariablen.

Da dieser Wert “andere” einen Key mit $k=1$ generiert, entfernen wir ihn vorsorglich, in dem wir ihn unter dem Menüpunkt “Set missing values to NA” durch ein **NA** ersetzen, siehe Abbildung 8.

Abbildung 8: Löschen der Kategorie “Andere” bei der Geschlechtsvariablen.

Zweitens rekodieren wir unter dem Menüpunkt “Convert numeric to factor” die Altersvariable `pi_ageinterview_22` von einer numerischen Variablen (Alter in Jahren) zu einer kategoriellen Variablen (5-Jahres-Altersklassen zwischen 50 und 105. Eingabe: 50,55,60,65,70,75,80,85,90,95,100,105), siehe Abbildung 9. Dies ermöglicht die Berücksichtigung des Alters in den Berechnungen zur k-Anonymity.

Abbildung 9: Rekodieren der Altersvariablen in eine Faktorvariable.

Abbildung 10 zeigt Verteilung bei der rekodierten Altersvariable.

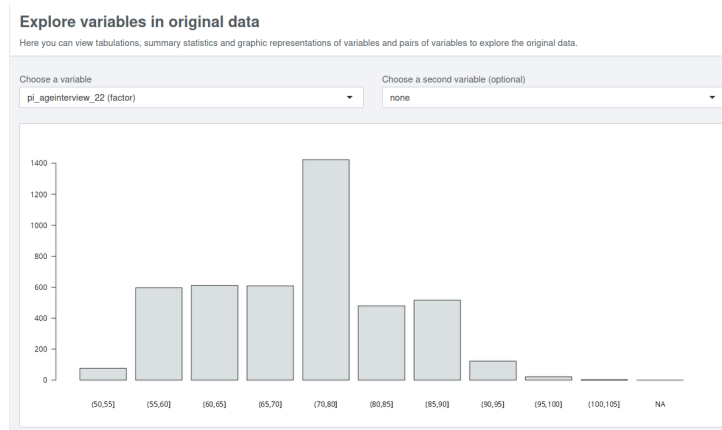


Abbildung 10: Verteilung der rekodierten Altersvariablen.

6.3 Anonymize

Unter dem Reiter “Anonymize” wird zuerst das Anonymisierungsproblem definiert, und anschliessend wird iterativ das Reidentifikationsrisiko bzw. k -Anonymity gemessen und Datenanpassungen gemacht, solange, bis $k=5$ eingehalten wird.

Abbildung 11 zeigt, wie für die reduzierten Pro Senectute Daten das Anonymisierungsproblem definiert wird. Insbesondere werden der Kanton, das Geschlecht und das Alter als Quasi-Identifiers bzw. Key Variables definiert. Die ID-Variable soll beim Erstellen der finalen, anonymisierten Daten weggelassen werden.

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 1

Variable name	Type	Key variables	Weight	Hierarchical identifier	PRAM	Delete	Number of levels	Number of missing
Respondent_ID	factor	☒ No ☐ Cat. ☐ Cont.	☐	☐	☐	☒	4457	0
canton_22	factor	☐ No ☒ Cat. ☐ Cont.	☐	☐	☐	☐	25	0
pl_sex_22	factor	☐ No ☒ Cat. ☐ Cont.	☐	☐	☐	☐	3	11
pi_ageinterview_22	factor	☐ No ☒ Cat. ☐ Cont.	☐	☐	☐	☐	11	0
ac_physact_22_freq	factor	☒ No ☐ Cat. ☐ Cont.	☐	☐	☐	☐	6	78
ac_cult_22_freq	factor	☒ No ☐ Cat. ☐ Cont.	☐	☐	☐	☐	6	169
ac_socact_22_freq	factor	☒ No ☐ Cat. ☐ Cont.	☐	☐	☐	☐	6	114
ac_concentrate_22_freq	factor	☒ No ☐ Cat. ☐ Cont.	☐	☐	☐	☐	6	112

[Setup SDC problem](#)

Abbildung 11: Anonymisierungsproblem definieren.

Nach Klick auf “Setup SDC problem” erscheint eine Übersichtsseite, siehe Abbildung 12.

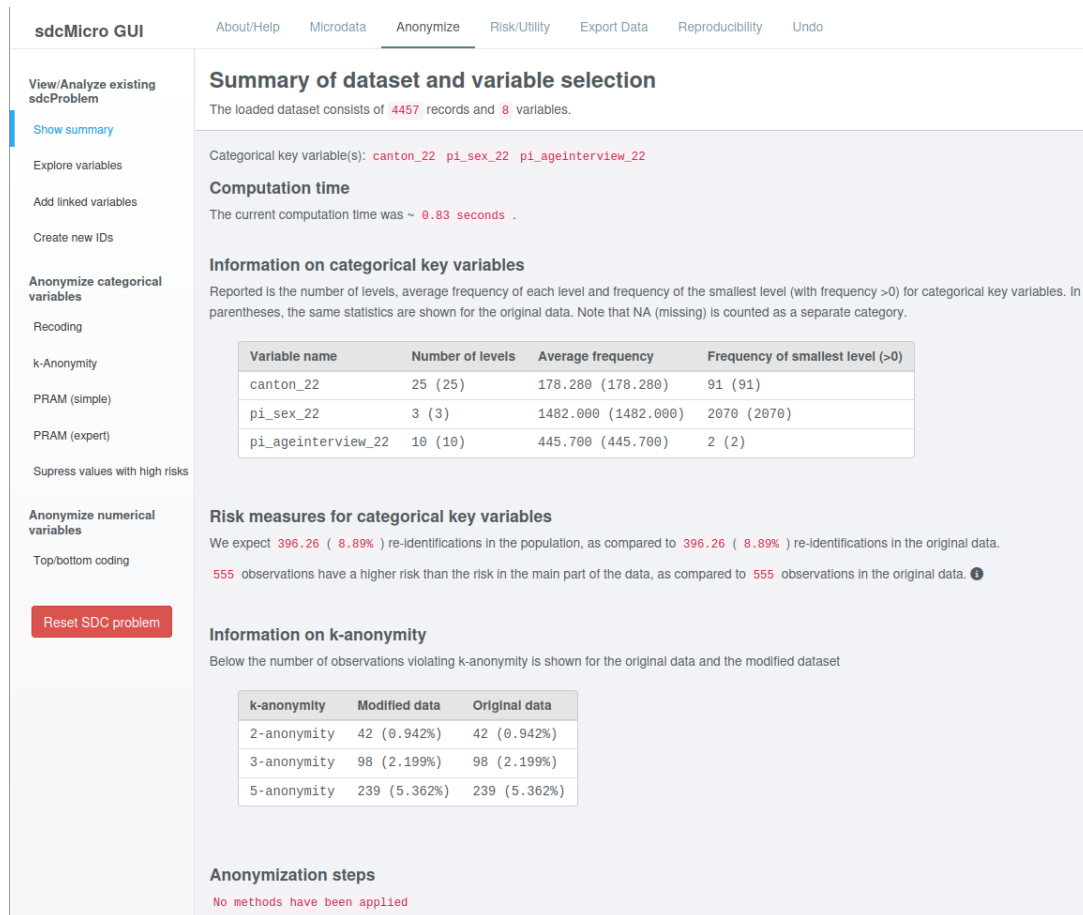


Abbildung 12: Erste Messung des Reidentifikationsrisikos

Darin sehen wir, dass aktuell $k=5$ bei 239 bzw. 5.36% der befragten Personen verletzt wird. Demnach muss mindestens einer der Quasi-Identifiers angepasst werden. Ein erstes Indiz, welcher der Quasi-Identifer angepasst werden soll, liefert der Absatz “Information on categorical key variables” in Abbildung 12. Wir konzentrieren uns vorerst auf das Alter, da dieses die Kategorie mit der tiefsten Häufigkeit aufweist.

6.3.1 Iteration 1

Kategorielle Variablen können unter dem Menüpunkt “Recoding” manuell angepasst bzw. vergrößert werden. Unter “Choose factor variable” wählen wir die Variable **pi_ageinterview_22**, siehe Abbildung 13. Wie ersichtlich ist, sind Altersklassen bis 55 und über 90 rar.

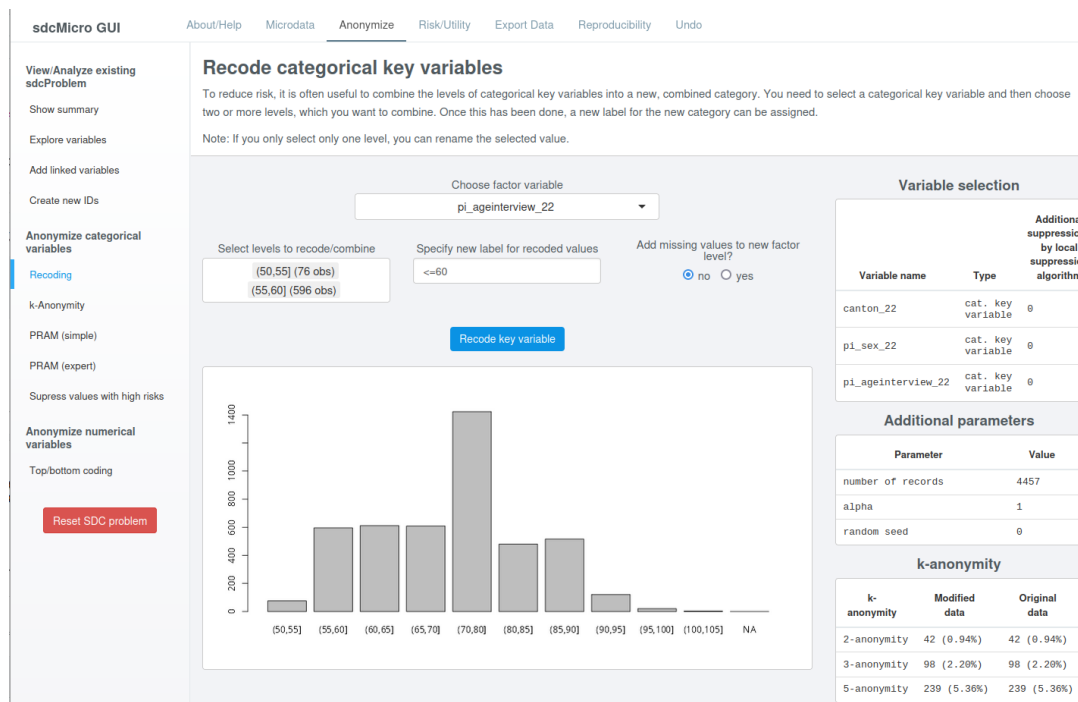


Abbildung 13: Zusammenführen der Alterskategorien bis 60 Jahre.

Abbildung 13 illustriert, wie die beiden Altersklassen (50,55] und (55,60] zu einer neuen Altersklasse ≤ 60 zusammengeführt werden. Analog werden auch die Kategorien (90,95], (95,100] und (100,105] zu einer neuen Altersklasse > 90 zusammengeführt. Abbildung 14 zeigt die Verteilung der rekodierten Altersvariablen.

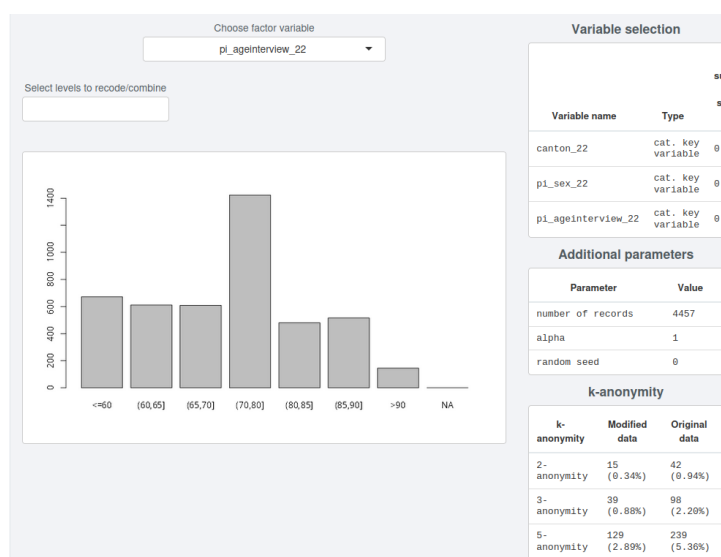


Abbildung 14: Verteilung der Altersvariablen nach dem rekodieren.

Wie aus der Tabelle “k-Anonymity” in Abbildung 14 ersichtlich ist, konnte die 5-Anonymity von 239 auf 129 bzw. 2.89% befragte Personen gesenkt werden.

6.3.2 Iteration 2

Unter “Show summary” ist nun ersichtlich, dass die Kantonsvariable mit 91 befragten Personen die kleinste Kategorie aufweist, siehe Abbildung 15.

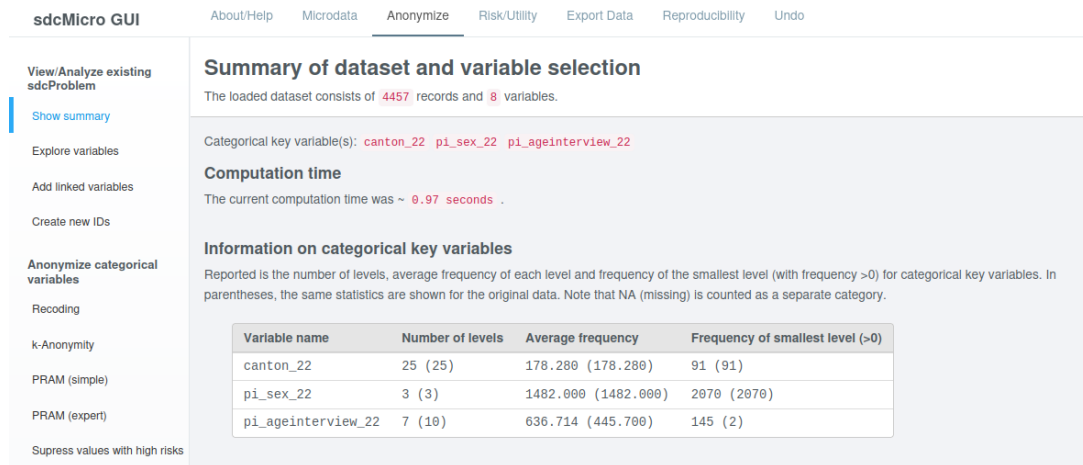


Abbildung 15: Verteilung der Altersvariablen nach dem rekodieren.

Wir wechseln nun wieder in den Menüpunkt “Recoding” um Kantone gezielt zusammenzuführen. Die Verteilung nach Kanton ist relativ ausgeglichen, siehe Abbildung 16. Dies hängt damit zusammen, dass die Stichprobe nach Kanton und Altersklassen stratifiziert ist.



Abbildung 16: Verteilung der Altersvariablen nach dem rekodieren.

Eine mögliche Strategie ist es, Schritt für Schritt 2 Kantone zusammenzuführen und zu beobachten, wie sich die Angaben zur k-Anonymity verhalten. Wir fokussieren hier auf eine radikalere Datenanpassung und führen die Kantone entsprechend den Grossregionen nach <https://de>.

[wikipedia.org/wiki/Grossregion_\(Schweiz\)](https://wikipedia.org/wiki/Grossregion_(Schweiz)) zusammen. Abbildung 17 zeigt, wie die drei Kantone Genf (GE), Waadt (VD) und Wallis (VS) zu einer gemeinsamen Kategorie GE_VD_VS zusammengeführt werden.

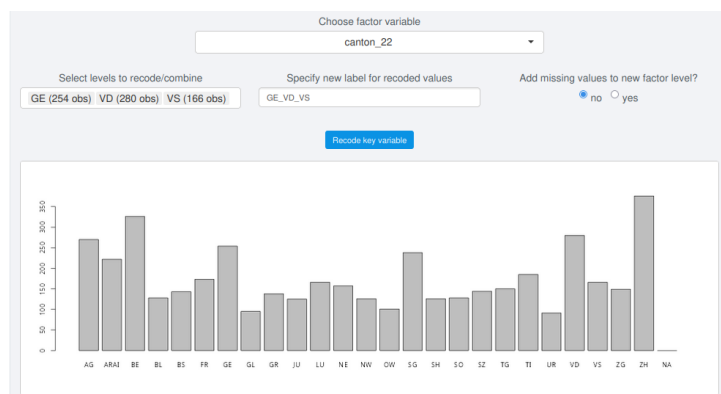


Abbildung 17: Verteilung der Altersvariablen nach dem rekodieren.

Abbildung 18 zeigt die Verteilung der Grossregionen und unten rechts die neuen Zahlen zur k-Anonymity. Die rarste Kategorie bei der Kantonsvariable ist der Kanton Tessin, der eine eigene Grossregion bildet. Die Datenanpassung hat dazu geführt, dass nur noch bei 7 bzw. 0.16% befragte Personen $k=5$ verletzt wird.

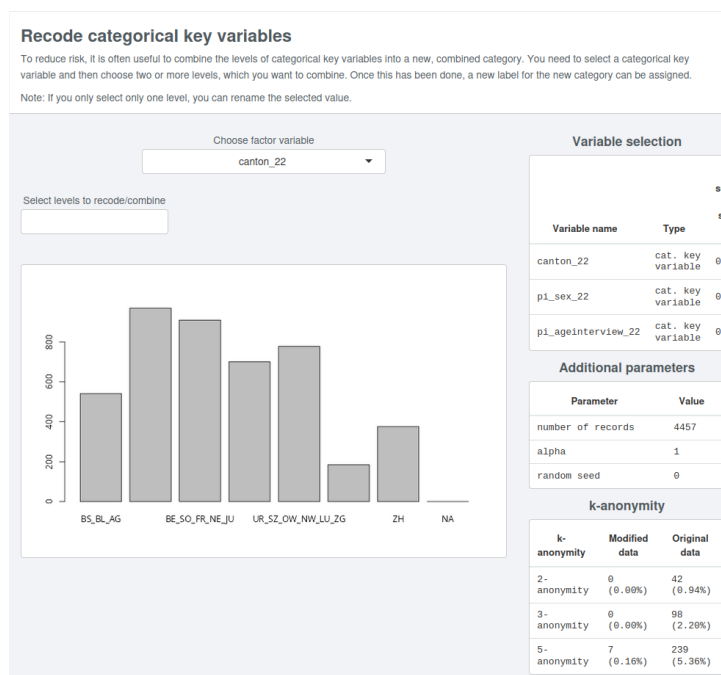


Abbildung 18: Verteilung der Altersvariablen nach dem rekodieren.

6.3.3 Iteration 3

Bei Iteration 3 wird ein zweites Mal die Altersvariable angepasst. Dabei wird die mit 145 befragten Personen kleinste Altersklasse >90 mit der nächsttieferen Klasse (85-90] zur neuen Klasse

>85 zusammengeführt. Abbildung 19 zeigt eine Zusammenfassung nach dieser Zusammenführung, inklusive Angaben zu den Datenanpassungen. $k=5$ wird nun von allen Keys eingehalten, und entsprechend werden keine weitere Anonymisierungsschritte unternommen.

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
Top/bottom coding
Reset SDC problem

Summary of dataset and variable selection

The loaded dataset consists of **4457** records and **8** variables.

Categorical key variable(s): **canton_22 pi_sex_22 pi_ageinterview_22**

Computation time
The current computation time was ~ **1.38 seconds**.

Information on categorical key variables
Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) 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 (>0)
canton_22	7 (25)	636.714 (178.280)	185 (91)
pi_sex_22	3 (3)	1482.000 (1482.000)	2070 (2070)
pi_ageinterview_22	6 (10)	742.833 (445.700)	490 (2)

Risk measures for categorical key variables
We expect **83.61** (**1.88%**) re-identifications in the population, as compared to **396.26** (**8.89%**) re-identifications in the original data.
17 observations have a higher risk than the risk in the main part of the data, as compared to **555** observations in the original data. ⓘ

Information on k-anonymity
Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	42 (0.942%)
3-anonymity	0 (0.000%)	98 (2.199%)
5-anonymity	0 (0.000%)	239 (5.362%)

Anonymization steps
Recoded "pi_ageinterview_22": "(50,55]", "(55,60]" to "<=60"
Recoded "pi_ageinterview_22": "(90,95]", "(95,100]", "(100,105]" to ">90"
Recoded "canton_22": "GE", "VD", "VS" to "GE_VD_VS"
Recoded "canton_22": "BE", "SO", "FR", "NE", "JU" to "BE_SO_FR_NE_JU"
Recoded "canton_22": "BS", "BL", "AG" to "BS_BL_AG"
Recoded "canton_22": "SG", "TG", "ARAI", "GL", "SH", "GR" to "SG_TG_ARAI_GL_SH_GR"
Recoded "canton_22": "UR", "SZ", "OW", "NW", "LU", "ZG" to "UR_SZ_OW_NW_LU_ZG"
Recoded "pi_ageinterview_22": "(85,90]", ">90" to ">85"

Abbildung 19: Zusammenfassung der anonymisierten Daten nach Iteration 3.

6.4 Export Data

Im Reiter “Export Data” können die anonymisierten Daten exportiert werden. Verschiedene Formate sind verfügbar. Beispielsweise kann wie in Abbildung 20 das CSV-Format gewählt werden, um die Daten mit Excel öffnen zu können. Die Daten werden in dem in Kapitel 6.1 festgelegten Verzeichnis abgelegt.

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

What do you want to export?
Anonymized Data
Anonymization Report

Export anonymized microdata

Select the file format to export the data to. If necessary, the order of the records can be randomized before exporting.

Show
10
entries

Search:

Respondent_ID	canton_22	pi_sex_22	pi_ageinterview_22	ac_physact_22_freq	ac_cult_22_freq	ac_socact_22_freq	ac_concentrate_22
10000	UR_SZ_OW_NW_LU_ZG	männlich	(65,70]	Mindestens einmal pro Woche	Mindestens einmal pro Jahr	Mindestens einmal pro Woche	Jeden Tag oder jeden Tag
10002	SG_TG_ARAI_GL_SH_GR	weiblich	(70,80]	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Jahr		Mindestens einmal pro Monat
10004	GE_VD_VS	weiblich	(80,85]	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Jahr	Mindestens einmal pro Woche	Nie
10005	UR_SZ_OW_NW_LU_ZG	weiblich	(70,80]	Mindestens einmal pro Woche	Mindestens einmal pro Monat	Mindestens einmal pro Monat	Jeden Tag oder jeden Tag
10009	ZH	männlich	(60,65]	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Monat	Mindestens einmal pro Woche	Mindestens einmal pro Woche
10011	SG_TG_ARAI_GL_SH_GR	männlich	(60,65]	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Monat	Nie	
10012	BS_BL_AG	männlich	(70,80]	Mindestens einmal pro Woche	Mindestens einmal pro Woche		Jeden Tag oder jeden Tag
10014	UR_SZ_OW_NW_LU_ZG	weiblich	(70,80]	Mindestens einmal pro Woche	Nie	Mindestens einmal pro Woche	Jeden Tag oder jeden Tag
10015	SG_TG_ARAI_GL_SH_GR	weiblich	(70,80]	Mindestens einmal pro Monat	Nie	Mindestens einmal pro Woche	Nie
10019	UR_SZ_OW_NW_LU_ZG	weiblich	>85	Jeden Tag oder fast jeden Tag	Mindestens einmal pro Jahr	Mindestens einmal pro Monat	Mindestens einmal pro Woche

Showing 1 to 10 of 4,457 entries

Previous
1
2
3
4
5
...
446
Next

Select file format for export

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

Set options for exporting a csv file

Include variable names in first row?

☒ Yes
☐ No

Field separator

☒ Comma
☐ Semicolon
☐ Tab

Decimal separator

☒ Decimal point
☐ Decimal comma

Randomize order of records

☒ No randomization
☐ Randomization at record level

Save dataset

Abbildung 20: Anonymisierte Daten exportieren.

6.5 Reproducibility

Im Reiter “Reproducibility” kann ein im Hintergrund generiertes R-Skript abgespeichert werden. Dieses R-Skript erlaubt es, die in der Webanwendung “per Klick” durchgeführte Analyse jederzeit in R zu reproduzieren. Das R-Skript wird in dem in Kapitel 6.1 festgelegten Verzeichnis abgelegt.

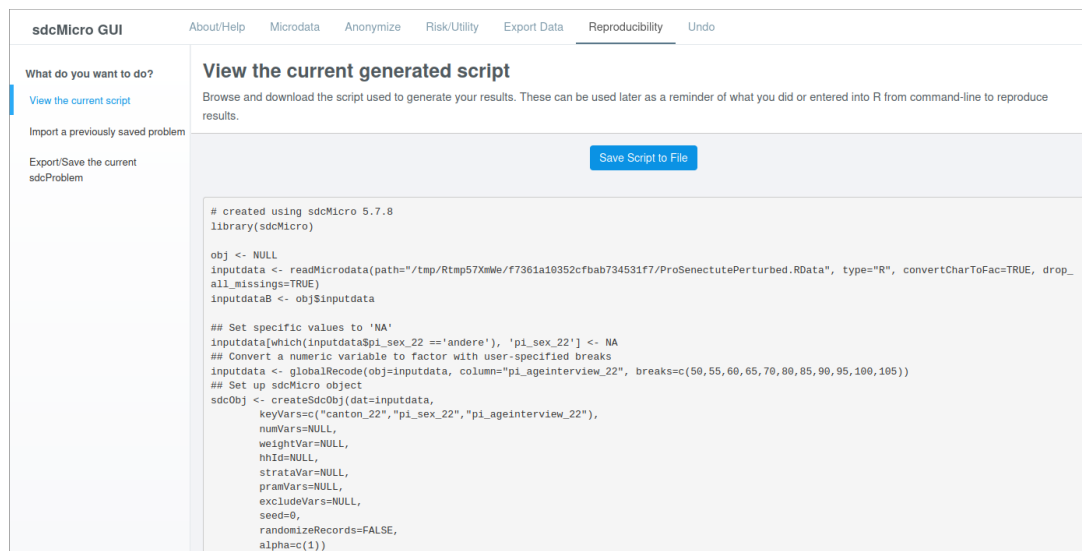


Abbildung 21: Automatisch generiertes R-Skript abspeichern.

Auch ist es möglich, im Menüpunkt “Export/Save the current sdcProblem” dass unter Kapitel 6.3 festgelegte Anonymisierungsproblem inklusive den Anpassungen etc. abzulegen, siehe Abbildung 22

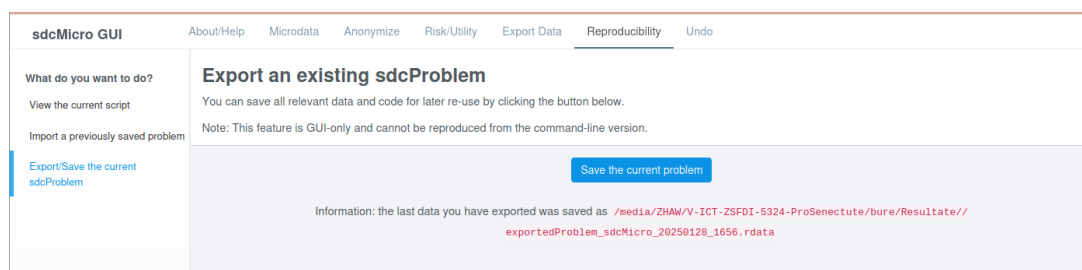


Abbildung 22: Anonymisierungsproblem exportieren

Nach Abschluss der Anonymisierung kann die Webanwendung im Reiter “About/Help” mittels Link “Stop the GUI” beendet werden.

Bemerkung: Die illustrierte Anonymisierung stellt nur eine mögliche Lösung dar, um $k=5$ zu erreichen. Die Lösung ist kaum optimal hinsichtlich dem Informationsverlust der Daten, d.h. wahrscheinlich gibt es alternative Lösungen, welche die Quasi-Identifizier weniger stark modifizieren, und trotzdem $k=5$ trotzdem einhalten. Andererseits ist die illustrierte Lösung unserer Meinung nach einfach, nachvollziehbar und erfüllt die $k=5$ Vorgabe.

7 Data Utility und Informationsverlust

Für eine allgemeine Einführung und Übersicht zu Data Utility und Informationsverlust verweisen wir auf Templ (2017), Kap. 5. Im Allgemeinen wird Informationsverlust gemessen, in dem man die ursprünglichen und finalen Daten mittels einer Distanzmetrik vergleicht oder Abweichungen ausgewählter Kennzahlen (bspw. durchschnittliche Einkünfte) betrachtet. In diesem

Abschnitt bewerten wir Data Utility und Informationsverlust lediglich aus einer qualitativen Sichtweise.

Offensichtlich sind die finalen Daten durch das Weglassen und Vergröbern von Variablen weniger informativ als die ursprünglichen Daten. Im Wesentlichen fand folgender Informationsverlust statt:

- Vergrößerung oder Entfernung von **Quasi-Identifiern**:
 - Weglassen vom Grossteil der Quasi-Identifizier zwecks Vereinfachung der Illustration
 - Altersklassen statt Alter in Jahren
 - Nur sehr grobe Angaben zum Wohnort (Grossregionen)
- Inhaltliche Variablen wurden zwecks Vereinfachung der Illustration und zwecks zur Verfügbarkeit machen der Illustrationsdaten grösstenteils weggelassen oder randomisiert

Die Autoren haben eine weitere Analyse durchgeführt, in welcher alle Variablen berücksichtigt wurden. Bei Interesse kann die dazugehörige Dokumentation zur Verfügung gestellt werden.

8 Diskussion und Fazit

Dieses Dokument soll Forschenden der ZHAW anhand einer Illustration mit realen Umfragedaten einen praxisbezogenen Zugang zur Thematik der Datenanonymisierung geben. Schritt für Schritt wurden sensible Ausgangsdaten modifiziert, so dass diese auf Anfrage weitergegeben oder gar publiziert werden könnten. Für die Anonymisierung der Umfragedaten wurden ausschliesslich einfache Kriterien zur Beurteilung des Reidentifikationsrisikos und einfache Anonymisierungsmethoden verwendet. Insbesondere wurden die Daten durch Weglassen oder Vergröbern von Variablen und weiteren einfachen Methoden anonymisiert, so dass diese $k=5$ Anonymity erfüllten. Die Illustrationsdaten sowie sämtliche Resultate werden als Anhang zu diesem Dokument zur Verfügung gestellt. Die Originaldaten können aus Datenschutzgründen nicht zur Verfügung gestellt werden.

Wie bereits in der Einleitung erwähnt sind die hier verwendeten Anonymisierungsmethoden nicht zwingend die beste Wahl. Die Literatur bietet viele komplexe Anonymisierungsmethoden wie Microaggregation, Rank Swapping oder Shuffling, die in manchen Anwendungen dienlicher sind als die hier illustrierten. Wenngleich mit den hier gezeigten Methoden viele Daten anonymisiert werden können, empfehlen wir den Lesenden den Deep Dive in die Literatur zum Thema, bspw. Templ (2017).

Die Illustration konzentriert sich auf ein spezifisches Anonymisierungsproblem. Sie umfasst bestimmt nicht alle Probleme, die den Lesenden bei der Anonymisierung eines eigenen Datensatzes begegnen können. Beispielsweise wurde hier ignoriert, dass Quasi-Identifizier manchmal numerisch sind. Auch gibt es Spezialprobleme bei Längsschnitt- oder raumbezogenen Daten oder Filterfragen. Zudem könnte man bei der Anonymisierung von Umfragedaten das Design inkl. Gewichte berücksichtigen. Auch könnten die Daten hinsichtlich weiterer Szenarien, wie beispielsweise dem Attribute-Disclosure-Szenario, anonymisiert werden.

Für Anregungen, Fragen oder Anfragen zu einer Unterstützung bei einer Anonymisierung steht die AG Datenschutz & Informationssicherheit der ZHAW gerne zur Verfügung (researchdata@zhaw.ch).

Finanzierung

Dieses Dokument wurde im Rahmen des ZSF Projekts “Integrating Discipline-Specific Know-How into Data Stewardship (DSembedded)” erstellt.

Dankessagungen

Herzlichen Dank an die Pro Senectute und Rainer Gabriel für das zur Verfügungstellen der Pro Senectute Daten. Herzlichen Dank an Simon von Rekum, Andi Führholz, Matthias Templ und Rainer Gabriel für Anregungen und Rückmeldungen zum Dokument.

Referenzen

Gabriel, Rainer, und Sonja Kubat. 2022. *Pro Senectute Altersmonitor: Altersarmut in der Schweiz 2022*. Teilbericht 1. Pro Senectute Schweiz.

Templ, Matthias. 2017. *Statistical Disclosure Control for Microdata*. Methods and Applications in R. Springer.

Liste mit mit Beispielen zu typischen Quasi-Identifiers

Im Folgenden einige Beispiele von typischen Quasi-Identifiers in alphabetischer Reihenfolge. Die Liste ist nicht abschliessend und es ist auch möglich, dass einige der genannten Variablen in spezifischen Kontexten keine Quasi-Identifizier sind.

- Arbeitsgeber
- Beruf
- Bezirk
- Bildungsniveau
- Akademischer Titel
- Ethnische Herkunft
- Geburtsdatum
- Geschlecht
- Hobbies
- Kanton
- Mitgliedschaften in Vereinen
- Position (z.B. Senior Manager)
- Postleitzahl
- Telefonnummer
- Wohnort
- Zivilstand

Anhänge

- Illustrationsdaten: ProSenectutePerturbed.RData
- Ergebnisse:
 - exportedData_sdcMicro_20250130_1022.csv: Anonymisierte Daten im CSV-Format
 - exportedScript_sdcMicro_20250130_1022.R: R-Script zur Reproduktion der Anonymisierung
 - exportedProblem_sdcMicro_20250130_1023.rdata: Anonymisierungsproblem