Anonymisierung von Aggregat-Daten

Schweizer Statistiktage 2019
Cham

Rolf Schenker
11. November 2019

Stadt Zürich
Statistik

Was ist Anonymisierung

*Die **Anonymisierung** ist das Verändern personenbezogener Daten derart, dass diese Daten nicht mehr einer Person zugeordnet werden können.*

Üblicherweise publizieren statistische Ämter Daten als Kreuztabellen

Es gibt aber auch Fälle, in denen Kreuztabellen Rückschlüsse auf Einzelpersonen ermöglichen. Dann müssen weitere Massnahmen ergriffen werden.

Beispiele, die eine Anonymisierung benötigen (1)

- Ausländeranteil für kleine Gebiete
  - In den meisten Fällen unproblematisch
  - Beträgt der Anteil aber 100%, können Rückschlüsse auf Einzelpersonen gezogen werden: Alle Personen, die dort wohnen, haben keinen Schweizer Pass.

- Einkommen der Personen in Klassen
  - Für das Gros der Personen kein Problem
  - Wenn einzelne sehr wohlhabende Personen in diesem Gebiet wohnen, können Rückschlüsse auf ihr Einkommen gezogen werden

Beispiele, die eine Anonymisierung benötigen (2)

Einkommensverteilung nach Steuertarif

Steuerbares Einkommen	Grundtarif	Verheiratetentarif
0	27	23
100 – 30 000	31	32
30 100 – 60 000	15	51
…		
1 000 100 – 1 200 000	0	0
1 200 100 – 1 400 000	0	0
1 400 100 – 1 600 000	1	0
1 600 100 und mehr	0	0

Beispiele, die eine Anonymisierung benötigen (3)

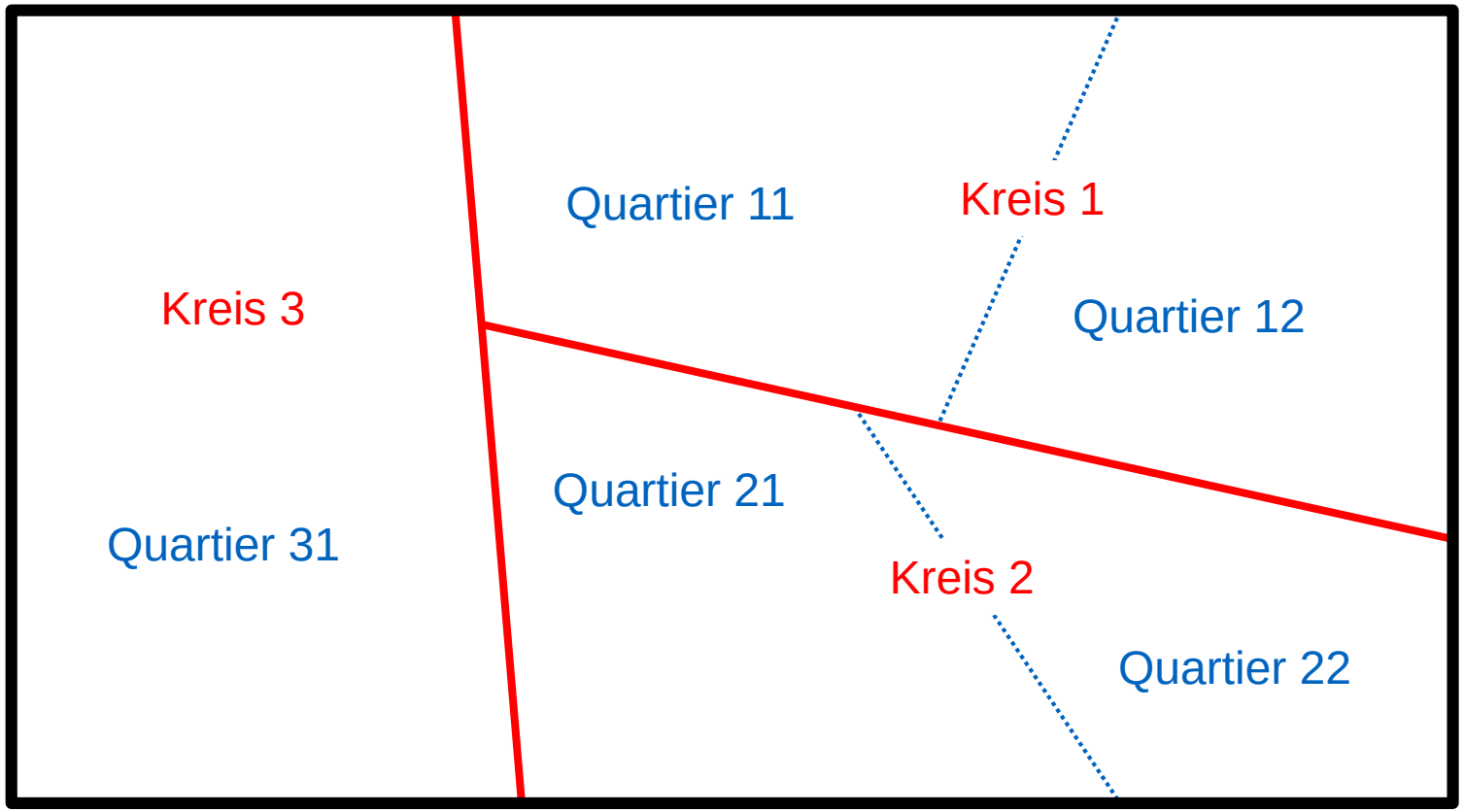
- Kleine Zahlen, die nicht präzise bekannt gegeben werden sollen
  - Anzahl Handänderungen von Grundstücken
  - Anzahl Betriebe nach Branche

Spezialfall: Hierarchische Klassifikation

- Berufe
- Branchen
- Räume

HIERARCHISCHE KLASSIFIKATIONEN

Hierarchische Klassifikation

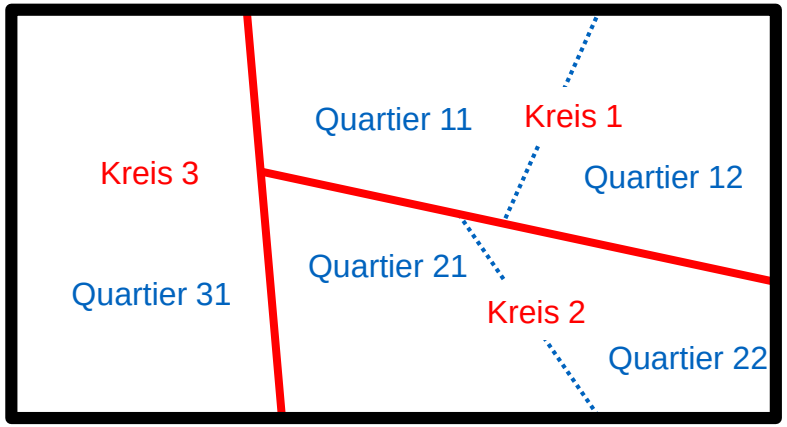


Ganze Stadt

Ein Beispiel

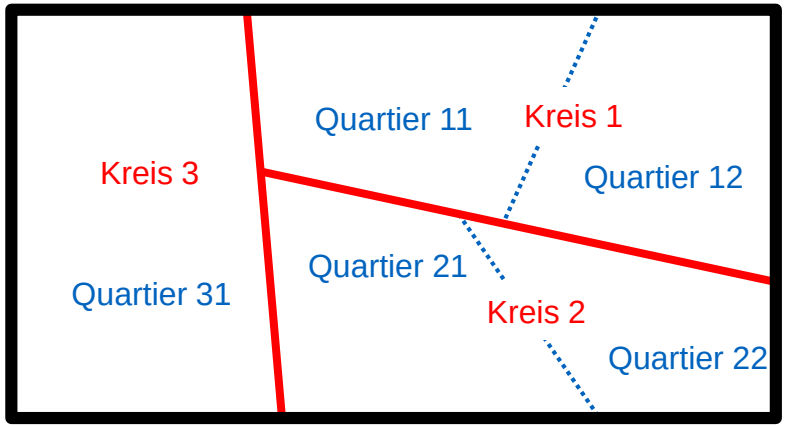
Gebiet	Wahrer Wert
Ganze Stadt	**28**
Kreis 1	**9**
Q 11	2
Q 12	7
Kreis 2	**2**
Q 21	2
Q 21	0
Kreis 3	**17**
Q 31	17

Regel: Werte zwischen 1 und 3 maskieren



Primäre Anonymisierung:
Maskieren auf der tiefsten Ebene

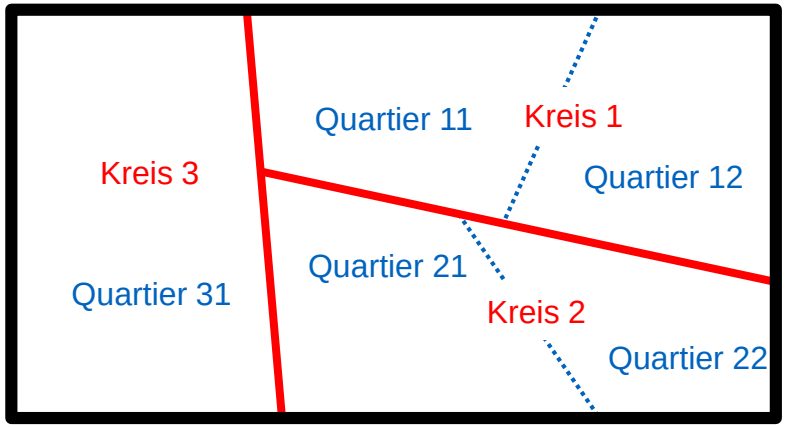
Ein Beispiel



Gebiet	Wahrer Wert	Primär	Sekundär	
			1. Schritt	2. Schritt
Ganze Stadt	**28**	**28**	**28**	**28**
Kreis 1	**9**	**9**	**9**	**9**
Q 11	2	1-3	1-3	1-3
Q 12	7	7	6-8	6-8
Kreis 2	**2**	**2**	**1-3**	**1-3**
Q 21	2	1-3	1-3	1-3
Q 21	0	0	0	0
Kreis 3	**17**	**17**	**17**	**16-18**
Q 31	17	17	17	17

Sekundäre Anonymisierung:
Sicherung der Anonymisierung mit Blick nach oben

Ein Beispiel



Gebiet	Wahrer Wert	Primär	Sekundär		Tertiär
			1. Schritt	2. Schritt	1. Schritt
Ganze Stadt	**28**	**28**	**28**	**28**	**28**
Kreis 1	**9**	**9**	**9**	**9**	**9**
Q 11	2	1-3	1-3	1-3	1-3
Q 12	7	7	6-8	6-8	6-8
Kreis 2	**2**	**2**	**1-3**	**1-3**	**1-3**
Q 21	2	1-3	1-3	1-3	1-3
Q 21	0	0	0	0	0
Kreis 3	**17**	**17**	**17**	**16-18**	**16-18**
Q 31	17	17	17	17	16-18

Tertiäre Anonymisierung:
Sicherung der Anonymisierung mit Blick nach unten

DAS VORGEHEN IM ÜBERBLICK

Vorgehen

1. Maskieren kleiner Werte
   - Auf der untersten Ebene

2. Sicherung der Anonymisierung mit Blick nach oben
   - Kann ein anonymisierter Wert auf der unteren Ebene mit dem Wert auf der oberen Ebene geknackt werden?
   - Ebenen werden von unten nach oben durchlaufen

3. Sicherung der Anonymisierung mit Blick nach unten
   - Kann ein anonymisierter Wert auf der oberen Ebene mit den Werten auf der unteren Ebene geknackt werden?
   - Ebenen werden von oben nach unten durchlaufen

Ziel des Vorgehens

- Einhaltung des Datenschutzes
  «Werte zwischen 1 und 3 als ‹1-3› darstellen»

- Sicherung der anonymisierten Werte durch Anonymisierung weiterer Werte

- Grundsatz: Je höher die Aggregationsebene,
  desto wertvoller die Information

  → Anonymisierung erfolgt «möglichst weit unten»

  → Information auf höherer Ebene möglichst unverändert publizieren

DER ALGORITHMUS

Der Algorithmus

- Der Anonymisierungs-Algorithmus
  - ist in SAS geschrieben
  - führt die Anonymisierung automatisch durch
  - vermeidet Fehler
  - verringert die benötigte Zeit deutlich

- Er erledigt das soeben vorgestellte Verfahren für beliebige Hierarchien

Inputs des Algorithmus

	Kreis	Quartier	Wert
Ganze Stadt	**0**	**0**	**28**
Kreis 1	**1**	**0**	**9**
Q11	1	1	2
Q12	1	2	7
Kreis 2	**2**	**0**	**2**
Q21	2	1	2
Q22	2	2	0
Kreis 3	**3**	**0**	**17**
Q31	3	1	17

Aufruf des Algorithmus

```
%anon_aggregatdaten(
  inputDaten = kleineStadt,
  inputVariable = Wert,
  outputDaten = kleineAnon,

  ebenen = quartier*kreis,
  minWert = 1,
  maxWert = 3,

  funktion = max
);
```

Datenset
Anonymisierungs-Variable
Datenset

Hierarchie-Stufen
Kleinster Wert, der maskiert wird
Grösster Wert, der maskiert wird

Auswahl der Region

Output des Algorithmus

	Kreis	Quartier	Wert	Anon	Phase	Grund
Ganze Stadt	**0**	**0**	**28**	**28**		
Kreis 1	**1**	**0**	**9**	**9**		
Q11	1	1	2	1-3	1	
Q12	1	2	7	6-8	2a	Sicherung Q11
Kreis 2	**2**	**0**	**2**	**1-3**	**2a**	**Sicherung Q21**
Q21	2	1	2	1-3	1	
Q22	2	2	0	0		
Kreis 3	**3**	**0**	**17**	**16-18**	**2b**	**Sicherung K2**
Q31	3	1	17	16-18	3a	Sicherung K2

OFFENE PUNKTE

Output des Algorithmus

- Von den anfangs aufgeführten Beispielen kann der Algorithmus einige lösen, andere sind noch offen
- Gelöste Beispiele
  - Anzahl Betriebe / Personen
- Ungelöste Beispiele
  - Anteile in hierarchischen Konstrukten
  - Einkommen der Personen in Klassen
- Grund-Problem: Wann muss etwas anonymisiert werden?
  - Statistik Stadt Zürich entwickelt derzeit einen Entscheidungsbaum / ein Scoring-System
  - Ziel ist, diese Frage für die meisten Situationen ohne Fachperson zu beantworten

Vielen Dank
für Ihre Aufmerksamkeit