4.4 Anonymisierung von Einzeldaten

4.4.1 Amtliche Statistik und Wissenschaft

Rat für Sozial- und Wirtschaftsdaten (RatSWD):

- 2004 vom Bundesministerium für Bildung und Forschung eingerichtet
- unabhängiges Gremium von empirisch arbeitenden Wissenschaftlern/-innen und Vertretern/-innen wichtiger Datenproduzenten
- Ziel: Verbesserung der Forschungsdateninfrastruktur für die empirische Forschung in den Sozial- und Wirtschaftswissenschaften
- Standardsetzung, Qualitätssicherung und weitere Entwicklung der Forschungsdatenzentren und Datenservicezentren

§ 16 Abs. 6 BStatG:

„Für die Durchführung wissenschaftlicher Vorhaben dürfen vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können [...].“

Anonymität von Einzeldaten (Mikrodaten) ist gegeben, wenn diese nicht dazu genutzt werden können, Informationen über die einzelnen statistischen Objekte zu erlangen.

Verschiedene Stufen der Anonymität:

- formale Anonymität: keine direkten Identifikationsmerkmale im Datensatz
- faktische Anonymität: Anonymität im Sinne des § 16 Abs. 6 BStatG
- absolute Anonymität: auch mit beliebig viel Zusatzwissen ist eine Reidentifikation Einzelner nicht möglich

Fiktives Datenbeispiel:

Vorname	Name	Stadtbezirk	Alter	Einkommen	Kfz-Marke
Marc	Böttcher	Sendling	33	2 650	Fiat
Daniel	Gruber	Maxvorstadt	26	890	Citröen
Maximilan	Held	Bogenhausen	46	3 200	BMW
Felix	Mayr	Schwabing-West	42	4 750	Porsche
Thomas	Pfeiffer	Au-Haidhausen	37	2 750	VW
Anton	Zander	Altstadt-Lehel	68	1 800	BMW

Zusatzinformationen:

- Mein Nachbar heißt Felix Mayr, ist 42, fährt einen Porsche und wir wohnen in Schwabing-West.
- Bei formaler Anonymisierung (Vorname und Name gelöscht): Aus einer öffentlichen Kfz-Statistik ist bekannt, dass es nur einen Fiat-Fahrer in Sendling gibt und mein Kollege, der in Sendling wohnt, fährt einen Fiat.

4.4.2 Anonymisierungsverfahren

Anonymisierungsverfahren können in zwei Gruppen eingeteilt werden:

- Verfahren zur Informationsreduktion
- Datenverändernde Verfahren

Verfahren zur Informationsreduktion

Merkmalsträgerbezogene Verfahren:

- Entfernen auffälliger Merkmalsträger
- Systematische Einschränkung der Grundgesamtheit
- (Sub-)Stichprobenziehung
  Dieses Verfahren wird u.a. bei der Anonymisierung der Mikrozensus-Daten eingesetzt.

Ausprägungsbezogene Verfahren:

- Löschung von seltenen Werten oder Merkmalskombinationen und Erzeugung von fehlenden Werten
- ggf. Ersetzung der fehlenden Werte

Merkmalsbezogene Verfahren:

- Beseitigung, Ersetzung oder Zusammenfassung von Merkmalen:
  - Unterdrückung einzelner Variablen
  - Ersetzen mehrerer Merkmale durch Linearkombination als neues Merkmal
  - Ersetzen mehrerer Merkmale durch Verhältniszahl als neues Merkmal
  - Indexzahl zu plausibler Basis anstelle der absoluten Werte
- Vergröberung von Merkmalsausprägungen:
  - Gruppierung von metrischen Merkmalen in Klassen
  - Rundung metrischer Werte
  - Zusammenfassung von Kategorien bei kategorialen Merkmalen

Fiktives Datenbeispiel:

Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
Sendling	3	82 650	500
Maxvorstadt	5	125 200	2 100
Bogenhausen	4	98 020	1 260
Schwabing-West	22	550 180	2 900
Au-Haidhausen	7	164 800	790
Altstadt-Lehel	4	108 450	1 100

- Stichprobe mit 6 von 30 Münchner Unternehmen einer Branche
- sensible Informationen sind hier Marketing-Ausgaben
- bekannt sind die Umsatzzahlen und der Standort der einzelnen Unternehmen

Fiktives Datenbeispiel: Merkmalsträgerbezogene Anonymisierung

Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
Sendling	3	82 650	500
Maxvorstadt	5	125 200	2 100
Bogenhausen	4	98 020	1 260
~~Schwabing-West~~	~~22~~	~~550 180~~	~~2 900~~
Au-Haidhausen	7	164 800	790
Altstadt-Lehel	4	108 450	1 100

Fiktives Datenbeispiel: Ausprägungsbezogene Anonymisierung

Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
Sendling	3	82 650	500
Maxvorstadt	5	125 200	2 100
Bogenhausen	4	98 020	1 260
NA	22	NA	2 900
Au-Haidhausen	7	164 800	790
Altstadt-Lehel	4	108 450	1 100

Fiktives Datenbeispiel: Merkmalsbezogene Anonymisierung

Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
~~Sendling~~	3	~~82 650~~	500
~~Maxvorstadt~~	5	~~125 200~~	2 100
~~Bogenhausen~~	4	~~98 020~~	1 260
~~Schwabing-West~~	22	~~550 180~~	2 900
~~Au-Haidhausen~~	7	~~164 800~~	790
~~Altstadt-Lehel~~	4	~~108 450~~	1 100

Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
München-Süd	3	$0 - 150\,000$	500
München-West	5	$0 - 150\,000$	2 100
München-Ost	4	$0 - 150\,000$	1 260
München-West	22	$> 150\,000$	2 900
München-Ost	7	$> 150\,000$	790
München-Zentrum	4	$0 - 150\,000$	1 100

Datenverändernde Verfahren

Swapping:

- Werte werden zwischen Merkmalsträgern zufällig vertauscht
- bei mehreren Merkmalen wird die Vertauschung für jedes Merkmal getrennt vorgenommen
- einfaches Data-Swapping: Merkmalsträger werden anhand ausgewählter kategorialer Merkmale gruppiert und die Werte der restlichen Merkmale werden innerhalb der Gruppen zufällig vertauscht
- Rank-Swapping: für jedes Merkmal werden die Werte der Größe nach sortiert und dann innerhalb festgelegter Nachbarschaftsbereiche zufällig getauscht
- bei Swapping bleiben die univariaten Verteilungen erhalten
- aber keine Zusammenhangsanalysen möglich, da sich gemeinsame Verteilung der Merkmale ändert

Fiktives Datenbeispiel: Data-Swapping

Rechtsform	Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
KG	Sendling	3	82 650	500
KG	Maxvorstadt	5	125 200	2 100
GmbH & Co. KG	Bogenhausen	4	98 020	1 260
GmbH & Co. KG	Schwabing-West	22	550 180	2 900
GmbH & Co. KG	Au-Haidhausen	7	164 800	790
KG	Altstadt-Lehel	4	108 450	1 100

- Gruppierung nach Rechtsform, zufällige Vertauschung der anderen Merkmalswerte

Rechtsform	Stadtbezirk	Mitarbeiter	Umsatz	Marketing-Ausgaben
KG	Altstadt-Lehel	4	108 450	2 100
KG	Sendling	3	82 650	1 100
GmbH & Co. KG	Schwabing-West	7	550 180	790
GmbH & Co. KG	Au-Haidhausen	4	164 800	1 260
GmbH & Co. KG	Bogenhausen	22	98 020	2 900
KG	Maxvorstadt	5	125 200	500

Mikroaggregation:

- Objekte werden zu Gruppen zusammengefasst und die Ursprungswerte jeweils durch das arithmetische Gruppenmittel ersetzt
- Gruppengröße mindestens drei Merkmalsträger
- zwei Typen nach der Bestimmung der Gruppen
  - deterministische Mikroaggregation
  - stochastische Mikroaggregation
- Erwartungswerte können korrekt geschätzt werden, Varianzen werden systematisch unterschätzt
- Zusammenhangsanalysen liefern unter Umständen verzerrte Ergebnisse

Deterministische Mikroaggregation:

- möglichst ähnliche Objekte zu Gruppen zusammenfassen
- gemeinsame Mikroaggregation:
  - nach einer Variablen
  - nach einer Hilfsvariablen
  - nach allen p metrischen Variablen: Bestimmung der Gruppen auf Basis der euklidischen Distanz im $\mathbb{R}^p$ für $\boldsymbol{x}_i, \boldsymbol{x}_k$ Datenvektoren von zwei Merkmalsträgern

$$\|\boldsymbol{x}_i - \boldsymbol{x}_k\|_2 = \sqrt{\sum_{j=1}^{p} (x_{i,j} - x_{k,j})^2}$$

- getrennte Mikroaggregation: Mikroaggregation wird für jedes Merkmal einzeln durchgeführt

Fiktives Datenbeispiel: gemeinsame Mikroaggregation

Mitarbeiter	Umsatz	Marketing-Ausgaben
3	82 650	500
5	125 200	2 100
4	98 020	1 260
22	550 180	2 900
7	164 800	790
4	108 450	1 100

- Mikroaggregation nach Variable Umsatz

Mitarbeiter	Umsatz	Marketing-Ausgaben
3	82 650	500
4	98 020	1 260
4	108 450	1 100
5	125 200	2 100
7	164 800	790
22	550 180	2 900

Mitarbeiter	Umsatz	Marketing-Ausgaben
3.67	96 373.33	953.33
3.67	96 373.33	953.33
3.67	96 373.33	953.33
11.33	280 060.00	1930.00
11.33	280 060.00	1930.00
11.33	280 060.00	1930.00

- Mikroaggregation nach Variable Marketing-Ausgaben

Mitarbeiter	Umsatz	Marketing-Ausgaben
3	82 650	500
7	164 800	790
4	108 450	1 100
4	98 020	1 260
5	125 200	2 100
22	550 180	2 900

Mitarbeiter	Umsatz	Marketing-Ausgaben
4.67	118 633.30	796.67
4.67	118 633.30	796.67
4.67	118 633.30	796.67
10.33	257 800.00	2 086.67
10.33	257 800.00	2 086.67
10.33	257 800.00	2 086.67

Stochastische Mikroaggregation:

- es werden zufällige Gruppen von Merkmalsträgern gebildet und die Werte durch die Gruppenmittelwerte ersetzt
- zufällige Gruppenzuteilung
- Bootstrap-Mikroaggregation

Weitere datenverändernde Anonymisierungsverfahren:

- Zufallsüberlagerung: Hinzufügen eines zufälligen Messfehlers
- Simulationsverfahren: Erzeugung synthetischer Datensätze auf Basis der gemeinsamen empirischen Verteilung