4 Datenproduktion: Erhebungsmethoden, Studiendesigns, Amtliche Statistik, Datenschutz und Anonymisierung

4.0 Vorbemerkungen

Literatur

- zu den Erhebunsmethoden und Studiendesigns (siehe: Literatur aus Kapitel 1.4f., 2 und 5), insb. Häder, M. (2010): *Empirische Sozialforschung*. 2. Auflage. Springer Verlag. Insbesondere Kapitel 6 und 7. Volltext-Download `http://www.ub.uni-muenchen.de/ausleihe-online/digitaler-zugriff/e-medien-login/index.html`, aufgerufen am 06.12.2016 im Rahmen des LRZ-Netzes.

- zur Amtlichen Statistik
  Rinne, H. (1996): *Wirtschafts- und Bevölkerungsstatistik*. 2. Auflage. Oldenbourg Verlag. Insbesondere Kapitel 2.
- Zu Anonymisierungsverfahren
  * von der Lippe, P. (1996): *Wirtschaftsstatistik*. 5. Auflage. Gustav Fischer Verlag. Insbesondere Kapitel 1.
  * Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. und Vorgrimler, D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. *Statistik und Wissenschaft 4*. Statistisches Bundesamt. Insbesondere Teil II.
  * Höhne, J. (2010): Verfahren zur Anonymisierung von Einzeldaten. *Statistik und Wissenschaft 16*. Statistisches Bundesamt. Insbesondere Einleitung – Kapitel 2.

4.1 Häufigste Erhebungsmethoden

Hier sind ganz kurz einige Methoden zur Erhebung von Daten aufgelistet, die bei wirtschafts- und sozialwissenschaftlichen Studien häufig verwendet werden.

- Beobachtung i.e.S.: unmittelbares, systematisches Erfassen der relevanten Sachverhalte
  * Feld versus Labor
  * teilnehmend versus nicht-teilnehmend
  * offen versus verdeckt

- Befragung: Datenerhebung anhand eines Fragebogens
  * klassisch face-to-face: persönlich vs. telefonisch vs. schriftlich (per Post)
  * Nutzung von moderner Technologie: CAPI - Computer Assisted Personal Interview vs. CATI - Computer Assisted Telefon-Interview vs. Online-Befragung
- Inhaltsanalyse: Auslesen von Daten aus Texten und anderen Quellen
- Nutzung prozessproduzierter Daten immer häufiger: Aufbereitung von Informationen, die nicht zu Forschungszwecken aufgezeichnet wurden
- Nicht-reaktive Verfahren i. e. S. („Spurensuche“)

4.2 Studiendesigns

Im Folgenden bezeichnet:

n: Stichprobenumfang

E: Erhebung der zu untersuchenden Merkmale

I_i: Untersuchungseinheit i, mit $i \in \{1, \ldots, n\}$, $n \in \mathbb{N}$

t_j: Erhebungszeitpunkt j, mit $j \in \{1, \ldots, T\}$, $T \in \mathbb{N}$

4.2.1 Querschnittsstudie

Erhebungszeitpunkt / Untersuchungseinheit	t_1	t_2	$\dots$	t_T
I_1	E			
I_2	E			
$\vdots$	$\vdots$			
I_n	E			

4.2.2 Zeitreihenstudie

Erhebungszeitpunkt / Untersuchungseinheit	t_1	t_2	$\ldots$	t_T
I_1	E	E	$\ldots$	E
I_2				
$\vdots$				
I_n				

4.2.3 Panel-Studie

Erhebungszeitpunkt / Untersuchungseinheit	t_1	t_2	$\dots$	t_T
I_1	E	E	$\dots$	E
I_2	E	E	$\dots$	E
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_n	E	E	$\dots$	E

4.2.4 Trendstudie

Erhebungszeitpunkt / Untersuchungseinheit	t_1	t_2	$\dots$	t_T
I_1	E			
I_2	E			
$\vdots$	$\vdots$			
I_{n_1}	E			
I_{n_1+1}		E		
I_{n_1+2}		E		
$\vdots$		$\vdots$		
I_{n_2}		E		
$\vdots$			$\ddots$	
$I_{n_{T-1}+1}$				E
$I_{n_{T-1}+2}$				E
$\vdots$				$\vdots$
I_{n_T}				E

4.2.5 Experimentelles Design

Im Folgenden bezeichnet:

R: Zuordnung der Untersuchungseinheiten zu den Versuchsgruppen per $Randomisierung$ (d.h. zufällig)

X_k: Einsatz des k-ten experimentellen $Stimulus$ bzw. der Behandlung k, mit $k \in \{1, \ldots, K\}$, $K \in \mathbb{N}$ *(„Treatments“)*

- Klassisches Experiment ($K = 1$):

t_0	t_1	t_2
R	X	E
	$-$	E

- Experiment mit Vorher-Nachher-Messung ($K = 1$):

t_0		t_1	t_2
R	E	X	E
	E	$-$	E

- Experiment mit verschiedenen Behandlungen:

t_0	t_1	t_2
R	X_1	E
	X_2	E

- Quasi-Experiment:

t_0		t_1	t_2
$-$	E	X_1	E
	E	X_2	E
	$\vdots$	$\vdots$	$\vdots$
	E	X_K	E

4.3 Amtliche Statistik

4.3.1 Träger und Ziele der amtlichen Statistik

Das Wort *Statistik* leitet sich vom lateinischen Wort *status* (u.a. Staat) und vom italienischen Wort *statista* (Staatmann) ab; (vgl. Kap. 1): „Lehre von den Staatsmerkwürdigkeiten“ (Achenwall).

Amtliche Statistik bezeichnet eine vom Staat angeordnete und von dessen Organen durchgeführte Statistik.

Informationelle Infrastruktur:
Umfassende, tief gegliederte, aktuelle Daten über die Bevölkerung, die Wirtschaft und die Gesellschaft sind Teil der notwendigen Infrastruktur eines Staates.
Rolle der amtlichen Statistik als „weitere Säule der Gewaltenteilung“.

„Die Statistik hat erhebliche Bedeutung für eine staatliche Politik, die den Prinzipien und Richtlinien des Grundgesetzes verpflichtet ist. Wenn die ökonomische und soziale Entwicklung nicht als unabänderliches Schicksal hingenommen, sondern als permanente Aufgabe verstanden werden soll, bedarf es einer umfassenden, kontinuierlichen sowie laufend aktualisierten Information über die wirtschaftlichen, ökologischen und sozialen Zusammenhänge. Erst die Kenntnis der relevanten Daten und die Möglichkeit, die durch sie vermittelten Informationen mit Hilfe der Chancen, die eine automatische Datenverarbeitung bietet, für die Statistik zu nutzen, schafft die für eine am Sozialstaatsprinzip orientierte staatliche Politik unentbehrliche Handlungsgrundlage.“Aus dem „Volkszählungsurteil“des Bundesverfassungsgerichts. Siehe z.B. `https://www.telemedicus.info/urteile/Datenschutzrecht/88-BVerfG-Az-1-BvR-209,-269,-362,-420,-440,-48483-Volkszaehlungsurteil.html`, aufgerufen am 06.12.2016.

Ohne amtliche Statistik

- würden Regierungen, Verwaltungen, Unternehmer etc. in den meisten Fällen ohne sachliche Fundierung handeln.
- wäre die Handlungsgrundlage weder für Handelnde noch für Öffentlichkeit nachprüfbar.

Bei den Trägern der amtlichen Statistik kann unterschieden werden:

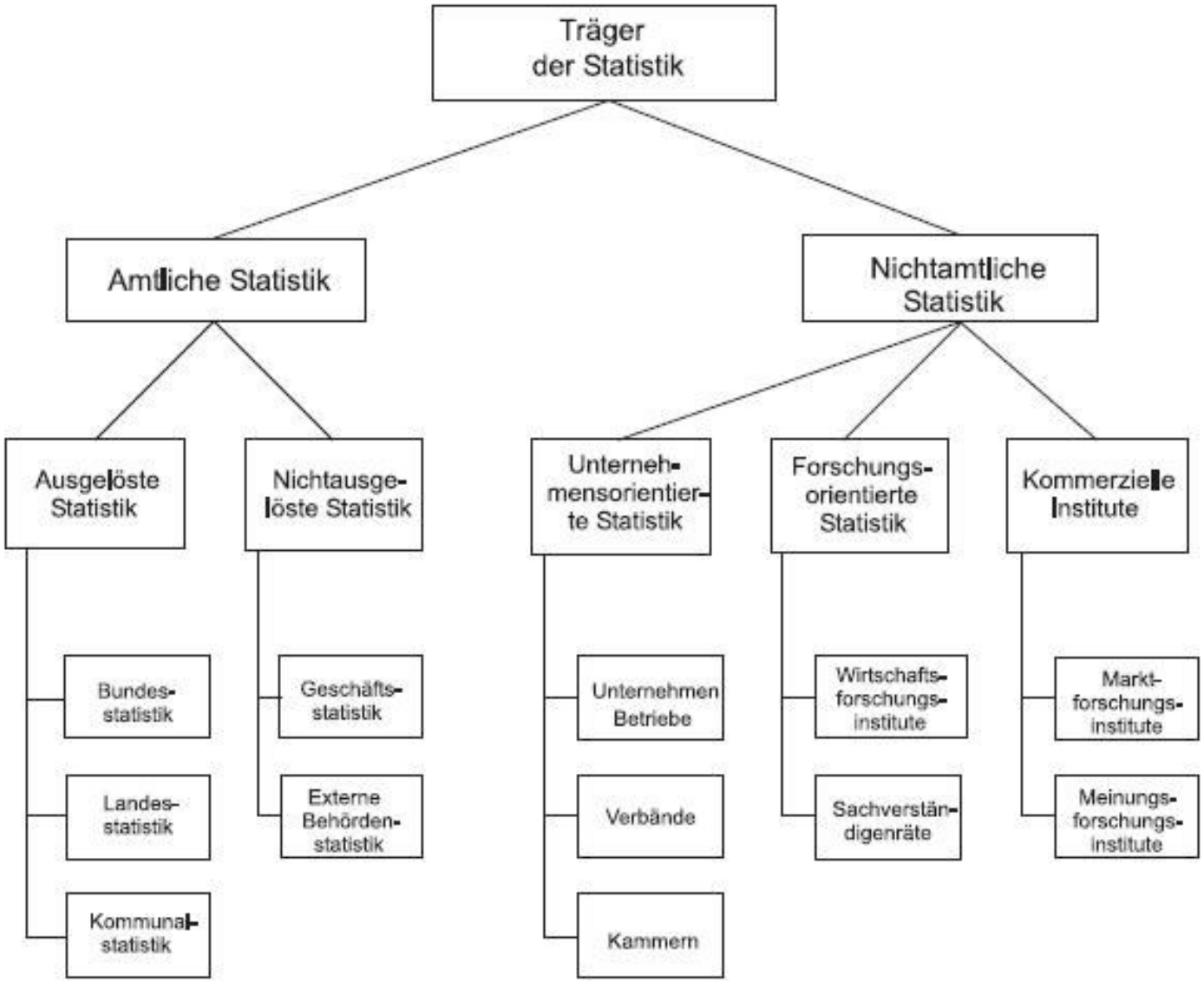
- ausgelöste Statistik: eigenständige Behörden, deren einziger Aufgabenbereich die Statistik ist
- nichtausgelöste Statistik: Abteilungen innerhalb von Behörden, die über die Geschäftsvorgänge Statistiken führen oder im Rahmen des Aufgabenbereiches der Behörde eigene Erhebungen durchführen

Nichtamtliche Statistik hat die auf individuelle Zwecke ausgerichtete Informationsbeschaffung zum Ziel.

Die Träger der nichtamtlichen Statistik lassen sich dem verfolgten Ziel nach einteilen:

- unternehmensorientierte Statistik
- forschungsorientierte Statistik
- kommerzielle Statistik

Amtliche und nichtamtliche Statistik in Deutschland (Quelle: Rinne (1996, S. 9))



Beispiele:

- Das Institut für Arbeitsmarkt- und Berufsforschung als eigenständige Dienststelle der Bundesagentur für Arbeit http://www.iab.de/de/ueberblick/gesetzlicher-auftrag.aspx, aufgerufen am 06.12.2016 IAB

- http://www.sachverstaendigenrat-wirtschaft.de/ziele.html, aufgerufen am 06.12.2016 Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung

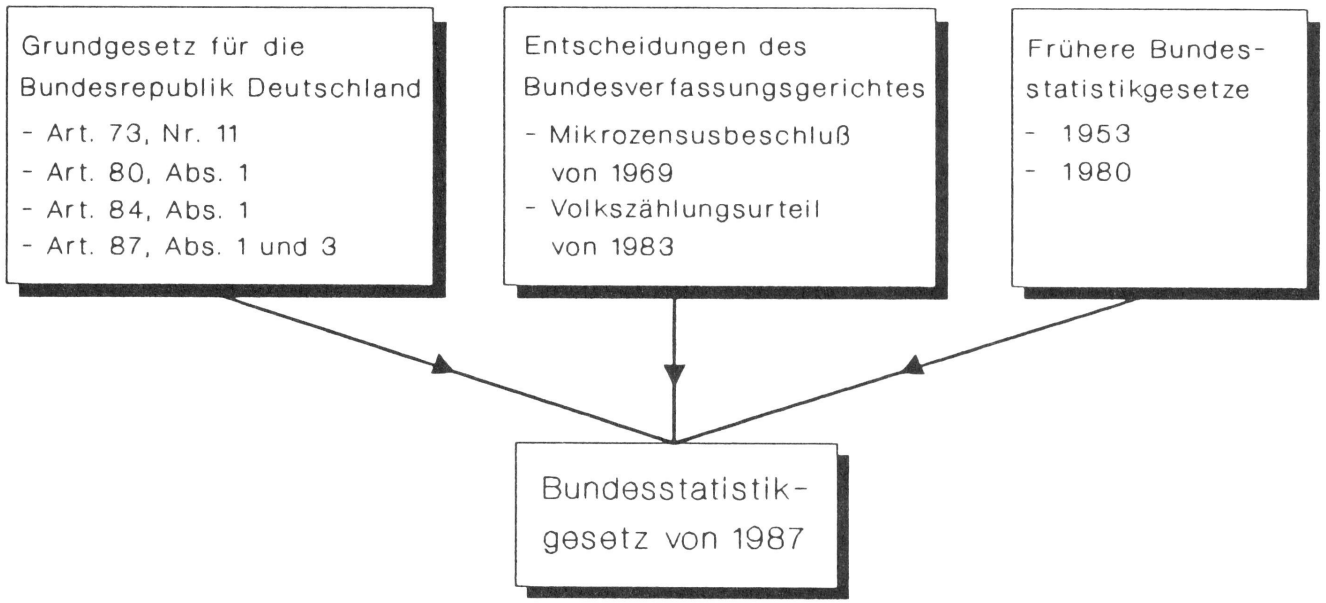
Beispiel

http://www.sachverstaendigenrat-wirtschaft.de/ziele.html, aufgerufen am 06.12.2016 Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung

4.3.2 Organisation der amtlichen Statistik in Deutschland

Rechtsgrundlagen für die Arbeit der Träger der amtlichen Statistik in Deutschland sind v.a. im Grundgesetz (GG, 1949) und im sog. Bundesstatistikgesetz (BStatG, 1987) verankert.

Rechtsgrundlagen der amtlichen Statistik (Quelle: Rinne (1996, S. 11), Kapitel 2)



Das Volkszählungsurteil von 1983 (`http://www.datenschutzbeauftragter-online.de/das-bundesdatens` `urteile-des-bverfg-zur-informationellen-selbstbestimmung/`, aufgerufen am 06.12.2016 BVerfGE 65, 1):

- Für das Frühjahr 1983 war ein Zensus durch Totalerhebung mit vielen Erhebungsmerkmalen geplant.
- Es gab einige Verfassungsbeschwerden von Bürgern gegen das Volkszählungsgesetz.
- Das Bundesverfassungsgericht erlässt nach der ersten mündlichen Verhandlung am 12. April 1983 eine einstweilige Anordnung, die Durchführung des Bundesgesetzes auszusetzen.

- Am 15. Dezember 1983 wird das Gesetz für verfassungswidrig erklärt. Begründung: Das Grundrecht auf informationelle Selbstbestimmung wird durch die im Volkszählungsgesetz vorgesehene Umsetzung eingeschränkt. Dieses Grundrecht wird abgeleitet aus:
  * Art. 2 Abs. 1 GG (Recht auf freie Entfaltung der Persönlichkeit)
  * Art. 1 Abs. 1 GG (Unantastbarkeit der Menschenwürde)

Auszug aus der Urteilsbegründung (`http://openjur.de/u/268440.html`, aufgerufen am 06.12.2016 BVerfGE 65, 1, Abschnitt C.II):

„Mit dem Recht auf informationelle Selbstbestimmung wären eine Gesellschaftsordnung und eine diese ermöglichende Rechtsordnung nicht vereinbar, in der Bürger nicht mehr wissen können, wer was wann und bei welcher Gelegenheit über sie weiß. Wer unsicher ist, ob abweichende Verhaltensweisen jederzeit notiert und als Information dauerhaft gespeichert, verwendet oder weitergegeben werden, wird versuchen, nicht durch solche Verhaltensweisen aufzufallen. [...] Dies würde nicht nur die individuellen Entfaltungschancen des Einzelnen beeinträchtigen, sondern auch das Gemeinwohl, weil Selbstbestimmung eine elementare Funktionsbedingung eines auf Handlungsfähigkeit und Mitwirkungsfähigkeit seiner Bürger begründeten freiheitlichen demokratischen Gemeinwesens ist.

Hieraus folgt: Freie Entfaltung der Persönlichkeit setzt unter den modernen Bedingungen der Datenverarbeitung den Schutz des Einzelnen gegen unbegrenzte Erhebung, Speicherung, Verwendung und Weitergabe seiner persönlichen Daten voraus. Dieser Schutz ist daher von dem Grundrecht des Art 2 Abs. 1 in Verbindung mit Art 1 Abs. 1 GG umfasst. Das Grundrecht gewährleistet insoweit die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen.“

Rechtsgrundlagen der amtlichen Statistik im `http://www.gesetze-im-internet.de/bundesrecht/gg/gesamt.pdf`, aufgerufen am 06.12.2016 Grundgesetz (1949):

- Art. 73 Nr. 11 GG: Der Bund hat die alleinige Gesetzgebungskompetenz über die Statistik für Bundeszwecke.
- Art. 80 Abs. 1 GG: Die Bundesregierung kann per Gesetz ermächtigt werden, entsprechende Rechtsverordnungen zu erlassen. ($\rightarrow$ BStatG)
- Art. 83 GG: Die Länder führen die Bundesstatistik als eigene Angelegenheit aus.
- Art. 84 Abs. 1 GG: Die Länder können für die Ausführung der Bundesstatistik eigene Behörden einrichten. ($\rightarrow$ Statistische Landesämter)
- Art. 87 Abs. 3 Satz 1 GG: Der Bund kann per Gesetz für die Bundesstatistik selbständige Bundesoberbehörden einrichten. ($\rightarrow$ BStatG: Statistisches Bundesamt)

Rechtsgrundlagen im `http://www.gesetze-im-internet.de/bundesrecht/bstatg_1987/gesamt.pdf`, aufgerufen am 06.12.2016 Gesetz über die Statistik für Bundeszwecke (1987):

§1 BStatG: Statistik für Bundeszwecke

- laufend Daten über Massenerscheinungen erheben, sammeln, aufbereiten, darstellen und analysieren
- Grundsätze der Neutralität, Objektivität und wissenschaftlicher Unabhängigkeit
- Die für die Bundesstatistik erhobenen Einzelangaben dienen ausschließlich den im BStatG oder in einer anderen Rechtsvorschrift festgelegten Zwecken.

§2 BStatG: Statistisches Bundesamt

§3 BStatG: Aufgaben des Statistischen Bundesamts, u.a.

- methodische und technische Vorbereitung sowie Weiterentwicklung von Bundesstatistiken und internationale Statistiken (v.a. EU)
  Zusammenstellung und Veröffentlichung der Statistiken für das Bundesgebiet
- Sammlung und Veröffentlichung statistischer Daten anderer Staaten und der EU
- Aufstellung der Volkswirtschaftlichen Gesamtrechnungen
- Beratung und Gutachtertätigkeit sowie vielfältige Erhebungs- und Aufbereitungsarbeit für andere Bundesbehörden

§4 BStatG: Statistischer Beirat

§17 BStatG: Aufklärungspflicht gegenüber den Befragten, um die Akzeptanz der Bundesstatistik zu fördern

Der Präsident des Statistischen Bundesamts ist i.d.R. auch Bundeswahlleiter.

Arten von Bundesstatistiken:

- Die EU kann nationale Statistiken anordnen.
- Fachlich zuständige Bundesministerien können Bundesstatistiken in Auftrag geben.
- zentrale Bundesstatistiken
- nach Entstehung (vgl. Kap. 1.5)
  * Primärstatistiken: eigens für den Zweck der Statistik erhobene Daten
  * Sekundärstatistiken: Erfassung von Daten aus Unterlagen, die zu anderen Zwecken angelegt wurden

Legalisierung der Erhebungen (§5 und §9 BStatG):

- „Legalitätsprinzip:“ Da amtliche statistische Erhebungen für die Befragten einen weitreichenden Eingriff darstellen können, bedarf jede Erhebung grundsätzlich einer eigenen Rechtgrundlage. (Siehe die `https://www.destatis.de/DE/Methoden/Rechtsgrundlagen/Rechtsgrundlagen.html`, aufgerufen am 06.12.2016 Sammlung von Rechtsgrundlagen des Statistischen Bundesamts.)
- Die Rechtsgrundlage ist grundsätzlich ein Gesetz, unter den Voraussetzungen des § 5 Abs. 2 BStatG kann es aber auch eine Rechtsverordnung sein.
- Rechtgrundlage legt fest: Erhebungsmerkmale, Hilfsmerkmale, Erhebungsart, Berichtszeitraum/-zeitpunkt, Periodizität und Kreis der zu Befragenden
- Nach §15 Abs. 1 Satz 1 BStatG muss ebenfalls geregelt sein, ob und in welchem Umfang bei der Erhebung Auskunftspflicht besteht.

Auskunftspflicht der Befragten (§15 BStatG):

- §15 Abs. 3 BStatG: Der Befragte muss wahrheitsgemäß, vollständig, fristgerecht und unentgeltlich die gewünschten Angaben machen.
- Die Qualität der amtlichen Statistiken beruht größtenteils auf der Auskunftspflicht.

Geheimhaltungspflicht der Bundesstatistik (§16 BStatG):

- §16 Abs. 1 BStatG: Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind grundsätzlich geheimzuhalten.
- Erhebungsphase
  primäre Geheimhaltung: betrifft Erhebungsbeauftragte
- Aufbereitungsphase
  interne Geheimhaltung: Hilfsmerkmale müssen nach Prüfung der Vollständigkeit und Schlüssigkeit der Daten gelöscht werden, damit eine spätere Zuordnung von Angaben zu Personen unmöglich wird.

- Publikationsphase
  externe Geheimhaltung: Die Weitergabe von Einzeldaten an externe Stellen oder kommunale Statistikämter ist nur unter strengen Auflagen möglich.

Bundesstatistik in der Öffentlichkeit:

- Das Dilemma der amtlichen Statistik in der Gesellschaft
- Aufklärungs- und Belehrungspflicht (§17 BStatG)
- Veröffentlichung von Bundesstatistiken durch `http://www.destatis.de/jetspeed/portal/cms/`, aufgerufen am 06.12.2016 Statistisches Bundesamt

4.4 Datenschutz in Deutschland

`http://isi-web.org/images/about/Declaration-EN2010.pdf`, aufgerufen am 06.12.2016 Berufskodex für Statistikerinnen und Statistiker des ISI (International Statistical Institute)

Ethical Principle Nr. 12: Protecting the Interests of Subjects

„Statisticians are obligated to protect subjects, individually and collectively, insofar as possible, against potentially harmful effects of participating. This responsibility is not absolved by consent or by the legal requirement to participate. The intrusive potential of some forms of statistical inquiry requires that they be undertaken only with great care, full justification of need, and notification of those involved. These inquiries should be based, as far as practicable, on the subjects' freely given, informed consent.
The identities and records of all subjects or respondents should be kept confidential. Appropriate measures should be utilized to prevent data from being released in a form that would allow a subject's or respondent's identity to be disclosed or inferred.“

Das Anliegen des gesetzlichen Datenschutzes ist es allgemein, Informationen vor Missbrauch bei ihrer Verarbeitung zu schützen.

Beim gesetzlichen Datenschutz wird unterschieden zwischen:

- Datenschutz im weiteren Sinne: Schutz aller Daten vor Missbrauch
- Datenschutz im engeren Sinne: Schutz personenbezogener Daten vor Missbrauch bei der Datenverarbeitung

Gesetze zum Datenschutz:

- BStatG: enthält Regelungen zum Datenschutz im Kontext der amtlichen Statistik
- `http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf`, aufgerufen am 06.12.2016 Bundesdatenschutzgesetz (BDSG, 1990): allgemeine Datenschutzbestimmungen
- diverse Landesdatenschutzgesetze

Zur Wahrung des Rechts werden https://www.bfdi.bund.de/bfdi_wiki/index.php/Datenschutzbeauftr aufgerufen am 06.12.2016 Datenschutzbeauftragte eingesetzt, ferner gibt es das https://www.bsi.bund.de/DE/Home/home_node.html, aufgerufen am 06.12.2016 Bundesamt für Sicherheit in der Informationstechnik.

Weitere Bestimmungen des BDSG:

- §4 Abs. 1 BDSG: Personenbezogene Daten dürfen nur verarbeitet werden, wenn es das BDSG, ein anderes Gesetz oder der Betroffene selbst erlauben.

  Grundsätzlich gilt ein Verbot mit Erlaubnisvorbehalt

- Eine Einwilligung kann nur dann die Grundlage für eine Erhebung, Verarbeitung oder Nutzung personenbezogener Daten sein, wenn die Voraussetzungen des §4a BDSG erfüllt sind:
  * Die Einwilligung bedarf der Schriftform, soweit nicht wegen besonderer Umstände eine andere Form angemessen ist.
  * Der Betroffene ist vorher über die Tragweite seiner Einwilligung aufzuklären: „informierte Entscheidung“ (z.B. über den Zweck der Erhebung, Verarbeitung oder Nutzung). Soweit nach den Umständen des Einzelfalles erforderlich oder auf Verlangen ist der Betroffene auch darüber zu informieren, was geschieht, wenn er nicht einwilligt (z.B. dass Ansprüche verloren gehen können).
  * Die Einwilligung muss auf der freien Entscheidung des Betroffenen beruhen; d.h. sie muss frei von Zwang sein. In diesem Zusammenhang ist auch zu berücksichtigen, ob sich der Betroffene in einer besonderen Situation (z.B. Arbeitsverhältnis) befindet, oder ob aufgrund einer faktischen Situation (z.B. Monopolstellung desjenigen, der die Einwilligung einholen will) ein Zwang für den Betroffenen besteht.

  Bei der Verarbeitung besonderer Arten personenbezogener Daten gem. §3 Abs. 9 BDSG (Angaben über die rassische und ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben) muss sich die Einwilligung ausdrücklich auf diese Daten beziehen.
- Behörden, Unternehmen und natürliche Personen dürfen personenbezogene Daten verwenden, um
  * ihre gesetzlichen Aufgaben zu erfüllen (§13 Abs. 1 BDSG)

* berechtigte privatwirtschaftliche Interessen zu wahren (§28 Abs. 1 BDSG)

Formen von Datenmissbrauch

- Datenmissbrauch bei der Speicherung
- Datenmissbrauch bei der Löschung
- Datenmissbrauch bei der Übermittlung von Daten

Für die Gewährleistung von Datenschutz ist ein umfassendes Datensicherungssystem unerlässlich, mit dem sicher gestellt werden kann, dass:

- Datenzugriff nur für Berechtigte möglich ist
- keine unzulässige Verarbeitung der Daten geschieht
- Daten bei der Verarbeitung nicht verfälscht werden
- Daten reproduzierbar sind

Grundsätzliches Problem: Es sind nicht die Daten isoliert zu sehen, sondern die Daten mit vorhandenen oder beschaffbaren Zusatzinformationen.

Wie weit darf Datenschutz gehen?

- Spannungsfeld zwischen GG, Datenschutz und statistischem Nutzen
- Die amtliche Statistik ist strengen Rechtsvorschriften unterworfen, welche aus „rein statistischer Sicht“ zunächst gelegentlich als Einschränkung empfunden werden.
- Datenschutz ist Grundrechtsschutz und die Wahrung der informationellen Selbstbestimmung eine Funktionsbedingung einer menschenwürdigen Informationsgesellschaft.
- Die Verletzung der Auskunftspflicht ist eine Ordnungswidrigkeit, eine Verletzung des Datenschutzes hingegen eine Staftat (§203 StGB und §43 BDSG).
- Jeder gibt Daten über sich oft unbewusst Preis, etwa beim Verwenden von Suchmaschinen, gratis E-Mail-Accounts, Bonuskarten etc. Diese Daten werden von den Betreibern genutzt, im harmlosesten Fall für gezielte Werbung.
- Daten können heute nahezu unbegrenzt gespeichert werden und auch in vielen Jahren erst genutzt werden.
- Infos: Virtuelles Datenschutzbüro

4.5 Datenforschungszentren; Anonymisierung von Mikrodaten

- Amtliche Statistik unterliegt grundsätzlich einer Geheimhaltungspflicht von Einzelangaben (§16 BStatG: Geheimhaltungspflicht der amtlichen Statistik)
  Ausnahme, wenn sie dem Betroffenen nicht zugeordnet werden können (§1 Abs.1 Nr. 4 BStatG, und Regelungen zur Datenweitergabe u.a. an die Forschung Abs. 6-10.)

- Die politischen Entscheidungsträger profitieren von einer wissenschaftlichen Erforschung der vorhandenen Daten, z.B. Erkenntnisse über die aktuellen Lebensverhältnisse in Deutschland (z.B. Datenreport `https://www.destatis.de/DE/Publikationen/Datenreport/Datenreport.html`, aufgerufen am 06.12.2016): ein Gemeinschaftsprojekt des Statistischen Bundesamtes (Destatis), des Wissenschaftszentrums Berlin für Sozialforschung (WZB) und des Deutschen Instituts für Wirtschaftsforschung (DIW)) oder Erkenntnisse über den demografischen Wandel (z.B. MPI für demografische Forschung in Rostock), die über die reine Beschreibung hinausgehen.
  Die Erforschung der Ursachen und Zusammenhänge ist ebenso wichtige Grundlage für politische Entscheidungen.
- Es gibt auch ein berechtigtes wissenschaftliches Interesse an den Einzeldaten. Das liegt darin begründet, dass diese Daten einerseits ein riesiges Analysepotential besitzen (auch besonders gute Qualität) und sich andererseits gegenüber Eigenerhebungen von wissenschaftlichen Einrichtungen durch einen

bedeutend größeren Erhebungsumfang auszeichnen. (siehe Höhne (2010))

- Es besteht zusätzlich der Grundsatz der Wissenschaftsfreiheit in Deutschland, der einen gesellschaftlich hohen Stellenwert genießt. (Die Forschung ist v.a. staatlich gefördert und wissenschaftlicher Fortschritt auch ein Leistungsmerkmal des Staates.)
- Aus diesen Gründen gibt es ein öffentliches Interesse daran, der Forschung aktuelle Daten, insbesondere auch Einzeldaten, zur Verfügung zu stellen.

4.5.1 „Amtliche Statistik und Wissenschaft“, Forschungsdatenzentren

`http://www.ratswd.de/`, aufgerufen am 06.12.2016 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

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

Anonymität ist in Einzeldaten (Mikrodaten) dann gegeben, wenn diese – auch zusammen mit zugänglicher Zusatzinformation – nicht zur Gewinnung von Informationen über die einzelnen statistischen Objekte (Personen, Unternehmen etc.) dienen können. Eine Gewinnung von Informationen erfolgt üblicherweise in zwei Schritten:
Der erste Schritt ist die eindeutige Zuordnung eines einzelnen Objektes zu einem Datensatz in der Mikrodatendatei. (D.h. man weiß, zu wem die Zeile im Datensatz gehört.)
Danach können dann aus diesem Mikrodatensatz alle vorhanden Informationen abgelesen werden. Befinden sich darunter noch Informationen, deren Kenntnis nicht schon für die Zuordnung erforderlich war, hat man einen Informationsgewinn. (siehe Höhne (2010))

Verschiedene Stufen der Anonymität:

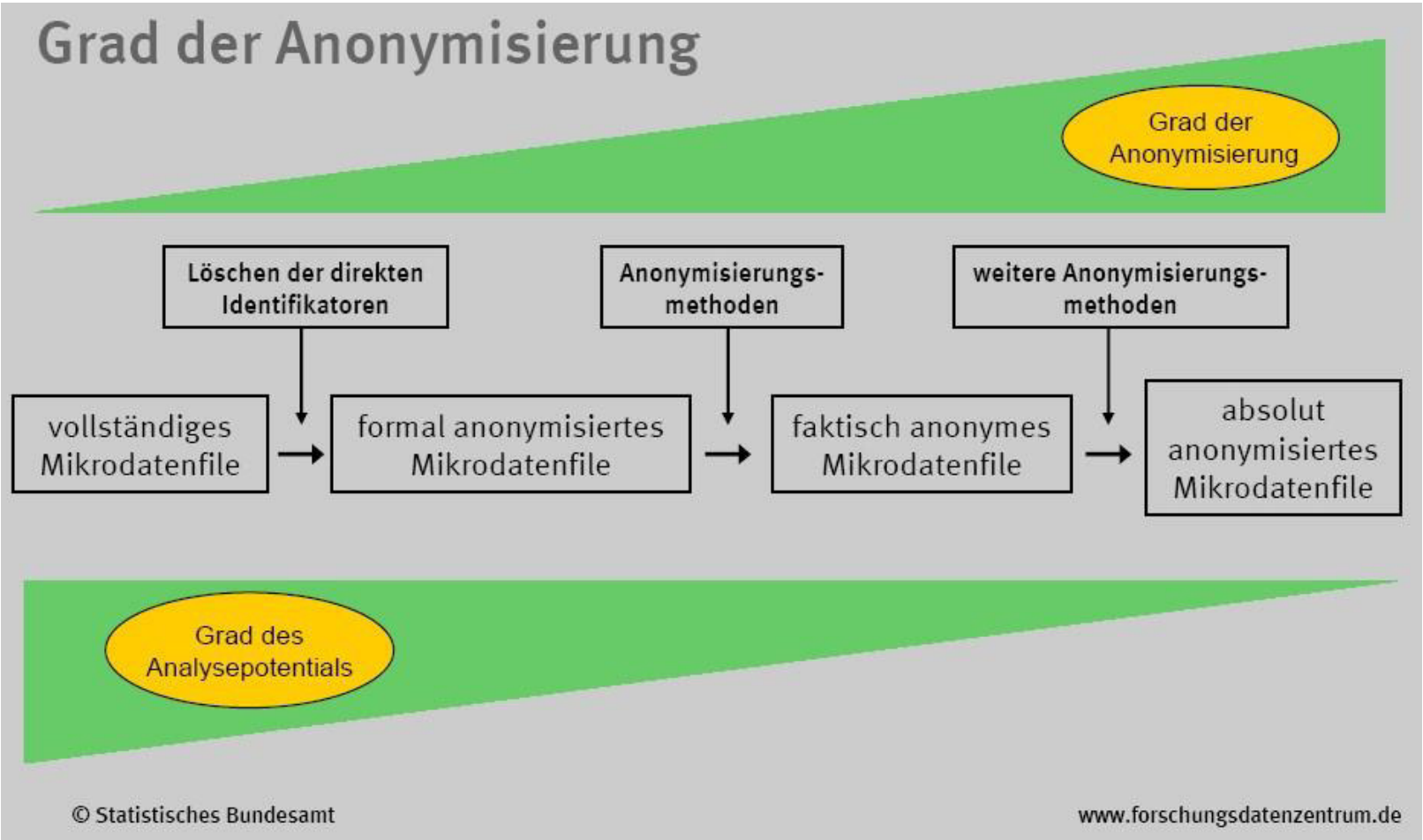
- formale Anonymität: keine direkten Identifikationsmerkmale im Datensatz
- faktische Anonymität: Anonymität im Sinne des §16 Abs. 6 BStatG ($\rightarrow$ scientific use Files)

  Identifikation nur mit unverhältnismäßigem Aufwand möglich
  §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 [...].“
und die Empfänger Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach Absatz 7 sind.

(7) Personen, die Einzelangaben nach Absatz 6 erhalten sollen, sind vor der Übermittlung zur Geheimhaltung zu verpflichten, soweit sie nicht Amtsträger oder für den öffentlichen Dienst besonders Verpflichtete sind. §1 Abs. 2, 3 und 4 Nr. 2 des Verpflichtungsgesetzes vom 2. März 1974 (BGBl. I S. 469, Artikel 42), das durch Gesetz vom 15. August 1974 (BGBl. I S. 1942) geändert worden ist, gilt entsprechend.

- absolute Anonymität: auch mit beliebig viel Zusatzwissen ist eine Reidentifikation Einzelner nicht möglich ($Campus\text{-}Files$)



Quelle: `www.empiwifo.uni-freiburg.de/lehre-teaching-1/Summer-term-10/Mat-Wirt-Sta/anonym`, aufgerufen am 06.12.2016.

Fiktives Datenbeispiel mit Betrieben:

Vorname	Name	Stadtbezirk	Alter	Einkommen	Kfz-Marke
Marc	Böttcher	Sendling	33	2 650	Austin-Mini
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 (Überschneidungsmerkmale)

4.5.2 Anonymisierungsverfahren

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. und Vorgrimler, D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. *Statistik und Wissenschaft 4*. Statistisches Bundesamt. Insbesondere Teil II.
bietet einen guten Überblick über Anonymisierungsverfahren

Anonymisierungsverfahren können in zwei Gruppen eingeteilt werden:

I) Verfahren zur Informationsreduktion
II) Datenverändernde Verfahren

I) Verfahren zur Informationsreduktion

Merkmalsträgerbezogene Verfahren:

- Entfernen auffälliger Merkmalsträger
- Systematische Einschränkung der Grundgesamtheit
- (Sub-)Stichprobenziehung
  Dieses Verfahren wird u.a. bei der Anonymisierung der `http://www.gesis.org/missy/metadata/MZ/`, aufgerufen am 06.12.2016 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

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 als Überschneidungsmerkmale 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

II) Datenverändernde Verfahren

Swapping:

- Werte werden zwischen Merkmalsträgern zufällig vertauscht
- bei mehreren sensiblen Merkmalen im Datensatz 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 für jedes Merkmal getrennt 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
- im Allgemeinen gilt: keine Zusammenhangsanalysen möglich, da sich gemeinsame Verteilung der Merkmale ändert (Verallgemeinerung: multivariates Swapping)

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	Nr.
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 bei den meisten Mikroaggregationsverfahren systematisch unterschätzt
- Zusammenhangsanalysen liefern typischerweise 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 in $\mathbb{R}^p$, definiert für $\boldsymbol{x}_i, \boldsymbol{x}_k$ Datenvektoren von zwei Merkmalsträgern als

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

- 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